A Sensible Sentimentalism

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Philosophy

by

Erick Jose Ramirez

Committee in charge:

Professor David Brink, Chair
Professor Dana Nelkin, Co-Chair
Professor Richard Arneson
Professor Christine Harris
Professor Piotr Winkielman

2012
The Dissertation of Erick Jose Ramirez is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

___________________________________________________________

___________________________________________________________

___________________________________________________________

___________________________________________________________

(Co-Chair)

___________________________________________________________

(Chair)

University of California, San Diego

2012
DEDICATION

To my parents, Luis and Leslie
TABLE OF CONTENTS

Signature Page ......................................................................................................................... iii

Dedication .............................................................................................................................. iv

Table of Contents .................................................................................................................. v

Acknowledgments ................................................................................................................ vi

Vita ........................................................................................................................................ vii

Abstract ............................................................................................................................... viii

Chapter 1: A Search for Sensible Sentimentalism .............................................................. 1

Chapter 2: On the Very Concept of Emotion .................................................................... 21

Chapter 3: Getting A Feel for Sentimentalism ................................................................. 59

Chapter 4: Rational Sentimentalism and Basic Emotions ............................................... 87

Chapter 5: The Constitutive Role and Uncodifiability ..................................................... 133

Chapter 6: Moral Responsibility, Reactive Attitudes and the Constitutive Claim .......... 172

References ............................................................................................................................ 224
ACKNOWLEDGEMENTS

Firstly to my advisers, David Brink, Dana Nelkin, and Dick Arneson. Your guidance, comments, and suggestions have been, and continue to be, invaluable. The dissertation would be twice and long and half as rigorous were it not for you. A special thank you to Justin D'Arms for accepting our invitation to share his work with the graduate students at UCSD and for helping to give birth to this project. To the members of the illustrious philosophy working group, Joyce Havstad, Adam Streed, and Tim Jankowiak. I could not have found a better group of friends or a better set of interlocutors. The dissertation has been improved immeasurably as a result of our discussions. To the rest of my friends in the department especially Lisa Damm, Charlie Kurth, Per Milam, Mitch Herschbach, Nat Jacobs, James Messina, Amanda Brovold, Eric Campbell, Amy Berg, Craig Agule, Cole Macke, Theron Pummer, and everyone else. I couldn't have asked for a more congenial group of philosophers to share my graduate experience with. To my friends and loved ones, Luis Ramirez, Leslie Delagneau, Margarita Levantovskaya, Tim Haag, Stacie Walsh, Nicolette Cook, my brother Luis, my sister Christian, my nephew Christopher, and the rest of my family. Thank you.

Chapter 5, in part, is being prepared for publication in *Philosophical Psychology*. The dissertation author was the primary author of this material.
VITA

2003        Bachelor of Arts, California State Polytechnic University Pomona
2007        Master of Arts, University of California, San Diego
2012        Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

“Critical Review: The Emotional Construction of Morals” by Jesse Prinz, forthcoming in Philosophical Psychology
ABSTRACT OF THE DISSERTATION

A Sensible Sentimentalism

by

Erick Jose Ramirez

Doctor of Philosophy in Philosophy

University of California, San Diego 2012

Professor David Brink, Chair
Professor Dana Nelkin, Co-Chair

The Sentimentalist argues that at least some evaluative properties, especially moral ones, are constituted by emotional responses. What is the best way to understand this claim? Against the received view of emotion, I develop a prototype theory according to which emotions are neither innate nor universal. Most contemporary Sentimentalists appeal to basic emotions to secure the objectivity of moral judgments. I argue, using the prototype theory of emotion, that emotions are not basic and that emotional judgments are best understood as subjective but not speaker relative and that a view of this kind can explain seemingly objective aspects of moral discourse. I close by challenging the claim that values are response-dependent by
criticizing contemporary arguments for response-dependence and proposing an empirical test for response-dependency claims within reactive attitude theories of moral responsibility.
Chapter 1

**A Search for Sensible Sentimentalism**

It should seem uncontroversial to say that our emotions are important to all of us. The United States Declaration of Independence names “the pursuit of Happiness” as one of three inalienable human rights.\(^1\) It would be hard to find a stronger statement of the value of our feelings. Our culture depicts agents not able to feel emotions as strange or seriously deprived and even as occasionally dangerous. Commander Data on Star Trek The Next Generation, for example, was an android built without the ability to feel emotions. Much of his narrative arc on the show revolved around his adopting the goal of being more human and for him this meant acquiring the capacity to feel. It also seems uncontroversial to say that our emotions play roles in our lives bigger than themselves. Here is what I mean: we do not simply care about feeling happy or feeling sad; we don't simply care about being jealous or being in love. We seem to care, quite a bit, about whether we *should* feel happy, sad, jealous, and so on and we publicly debate about what standards we should use to judge how we should feel.\(^2\) This concern with how we should feel is important but vague. There seem to be a lot of reasons that could lead us to think we should feel one way or another. It might be rude to feel happy about getting a job in front of your unemployed friends, it might not be in your interest to feel disgusted by the meal your well intentioned partner has made and it might seem immoral to laugh at offensive jokes.

---

2. As just one example consider the following headline which appeared shortly after President Barack Obama's announcement that Osama bin Laden had been killed by United States special forces members: “Osama bin Laden's death: How should we feel?” by Elizabeth Landau and Madison Park published electronically by CNN.
Considerations of etiquette, self-interest, and morality all seem to have a say in telling us how we *should* or *should not* feel. Though these are common reasons we give ourselves for directing our emotions there also seem to be distinctly emotional reasons for feeling. Consider landing a job. It might be rude to be happy in front of your unemployed friends. It might hurt them to be reminded of their situation and they might be jealous at your success. These reasons do suggest that maybe we ought not feel too happy in front of our friends if we get that job. On the other hand there seems something distinctly right about feeling happy. Landing a job, especially one that you have been hoping to get, is a good thing and we think that we should feel happy when good things happen to us. There seem to be some reasons for feeling that are expressly about our emotions as opposed to about politeness, our self interest, or morality. Some philosophers argue that these reasons are special. They claim that these kinds of reasons not only play very specific roles for directing how we should feel but also that they tell us something about the way values work. Values, at least some values, are thought to be dependent on our sentiments, our emotions, in a fundamental sense.

My focus in this dissertation is on these emotions, the values associated with these emotions, and the kinds of rules we appeal to when we regulate them. In particular I want to focus on the kinds of emotions and rules that we think work in the distinct way mentioned above. Emotions like outrage, indignation, and resentment seem especially tied to moral judgments or moral behavior and if there are emotion specific reasons for telling us how we *should* feel then there might be emotion specific
roles for our moral judgments also. It seems a truism to say, as I did at the outset, that our emotional responses matter to us. They can be used to reaffirm or challenge our existing values but they can also make us aware of values we may not have known that we held in the first place. For example if we do not quite *feel* content with the shiny new gadget we just bought and that we thought would make us happy, what might the lack of an expected emotion tell us? In this sense the lack of an expected emotion is informative. We might learn that we do not actually value new gadgets so much as something else about them, for example the envy they might produce in others or the freedom they might give us to spend time with friends and family. In these cases we normally turn, again, to our emotions to figure out what the answer might be. But to say much more than the fact that emotions matter to us requires moving beyond casual observations or platitudes and into a more serious examination of our emotions themselves.

Philosophers have long thought that emotions must be accounted for in a complete theory of morality. Some have thought that this meant only that we must be certain not to allow our emotions to guide our moral judgments. Others have argued in the opposite direction. These philosophers argue that the meaning of a moral judgment can only be found by understanding the emotions that our moral judgments express. I focus my attention on this latter set of moral theories. I have already mentioned one of these theories in passing. Philosophers who argue that there are distinct emotional reasons to feel, for example, think that the values associated with feelings like fear,
amusement, disgust, love, resentment, and many others each carve out a distinctly emotional set of values. These theories, call them Sentimentalist moral theories because of how they connect morality with our emotion, are a diverse lot. Although Sentimentalism will occupy much of my attention for several chapters of this book saying much more about them now would get ahead of the argument.

Sentimentalism, as I have defined it thus far, is committed to emotions playing an important role in our moral judgments or moral values. This implies that we have a good sense of just what it is that emotions are in the first place. This question is much more complicated than it might first appear. Think about the last emotional episode you remember feeling. Odds are that it was not very long ago. Now think back more carefully to the emotion itself. If you are like most people your emotional episode had many different parts or components to it. Most obvious is the feeling of the episode. It can be positive, negative, energetic, listless or all manner of states in between. There were also other components. You likely remember thinking specific things and making judgments about the objects and people around, you may even have acted as a result of these feelings and thoughts. All of these components are typical of an emotional episode but it is harder to say which, if any, of these components is the emotion.

If emotions are to serve some important role in moral judgment then we need to give an account of what an emotion is such that it can function in that important role. A theory of emotion then is foundational for Sentimentalists and it is for this reason that assessing theories of emotion will be my first order of business. A
Sentimentalist theory of morality is only as sensible as its theory of emotion.

1.1 A Theory of Emotion

Philosophers do not have a monopoly on theorizing about the emotions. Psychologists and neuroscientists also study them. The second chapter focuses on incorporating various contributions made in the multi-disciplinary study of emotion. It is often the case that parallel theories of emotion are developed in philosophy and psychology and that data from the neurosciences can play an important role in motivating one side of a debate over another. This kind of interdisciplinary approach to the study of emotion helps sheds light on what it is that emotions could be.

Part of my aim here will be to expose various points of disagreement within different theories of emotion. For example many disagree on where to locate the emotion in an emotional experience. Are emotions feelings? Thoughts? Actions? None of these or all of them? There is disagreement on where to draw the boundary between an emotion proper and emotional antecedents or effects. Furthermore, there is disagreement on the components of our emotional experience. Must emotions contain specific kinds of thoughts or judgments in order to 'count' as being an emotion of the right sort? Must resentment, for example, include the judgment that I have been wronged or can I resent someone without making judgments? Theories of emotion falling within the 'cognitivist' tradition argue that some thoughts, beliefs, or judgments must always be present. Non-cognitivists argue the opposite. While thoughts, beliefs, or judgments may be present during an emotional experience they are not necessary.\(^2\)

\(^2\) The cognitivism/non-cognitivism terminology here is entirely separable from debates about
A related issue: there are many components present during an emotional episode.

There is disagreement among theorists about whether any of these specific components should be seen as being necessary or jointly sufficient for an episode to be seen as an emotional episode. A non-cognitivist 'feeling' theory, for example, would suggest that only a distinct emotionally specific 'feel' needs to be present in order for an episode to count as an instance of an emotional type. Other theorists may focus on behavior or thoughts or visceral changes. Still other theorists argue that some set of these components may be necessary but only jointly sufficient. These theories are what some have called 'emotional hybrid' theories and these hybrid theories represent the most common way in which theorists of emotion have analyzed our feelings. One recently popularized hybrid of this kind argues that emotions should be identified and differentiated via our perception of different physiological changes and that different physiological changes come to represent different sets of concerns. This combines a non-cognitive physiological theory of emotion with a perceptual one and ties them with an evolutionary account of the origin of affect in order to explain the content of emotional responses.\(^3\)

There is another important question here on which emotion theorists divide. I

cognitivism and non-cognitivism in meta-ethics though they may intersect. Emotional cognitivism, as I explain, is a statement about the content of an emotional concept or experience: that it necessarily requires propositional attitudes. Meta-ethical cognitivism is a statement about the content of our moral utterances: that they are propositional in nature (and hence can take a truth value). Although these are separable claims the relationship between the two is fairly clear: if Sentimentalists believe our moral judgments express our emotions and they believe our emotions are cognitive, they have a cognitive theory of emotion, then they are likely to take have a cognitivist understanding of moral judgments as well.

have already alluded to the issue of how to structure emotional episodes. In that case the question was what parts of the episode to include as 'the emotion' and which to include as its causes or effects. There is a second issue here: how are emotional episodes structured? Jesse Prinz has labeled this the Problem of Plenty. Emotional episodes, whether cognitive or non-cognitive, are complex events. Is each emotion definable in terms of a set of necessary and sufficient conditions or are emotions less orderly than this? Most theorists side with the former view. Emotions, it is typically believed, are highly stereotyped events and each emotion can be identified with its distinct causes, beliefs (if cognitive), 'feel,' and the behavioral tendencies associated with it. I argue that these views, though initially attractive, are ultimately not the best way to explain emotional episodes as we experience them or to explain important cross-cultural differences between conceptions of emotions.

I offer a theory of emotion that addresses much of these disagreements. I argue that one family of theory, Psychological Constructivism, is best positioned to address the kinds of questions about our emotions that I have been raising. Psychological Constructivists believe that emotions are best thought of as not having necessary or jointly sufficient components. Instead emotions should be seen as structured by what are sometimes called 'paradigm scenarios' or prototypes or scripts. These paradigm scenarios are themselves built up out of a variety of different components only some of which are shared by all emotions. Each paradigm scenario, each prototype, fits within a certain cultural and historical context. Not all cultures will develop the same
paradigm scenarios and hence not all cultures will have access to the same set of emotions. Viewing emotions in this way, I argue, helps to make sense of many of our emotional experiences in ways other theories can not.

1.2 Basic Emotions

There is another important question about emotions. Are any emotions 'basic'?

Many philosophers and psychologists believe that there are basic emotions. Basic, in this sense, implies that all humans share the same emotions because emotions are the product of a shared evolutionary heritage. A theory of emotion where at least some emotions are basic is committed to the view that emotions are distinct. This means that there are unique differences or markers for each type of emotion. These views are also committed to the idea that these basic emotions are nearly universally shared at least in part because they are heritable. Some emotions have been though to qualify as basic in this sense. Happiness, sadness, fear, disgust, anger, shame, and guilt have often taken to be heritable and distinct in ways that make them basic. This view has attracted Sentimentalist philosophers.

Sentimentalists appeal to basic emotions in order solve longstanding theoretical problems with explaining some aspects of moral discourse. The third chapter is devoted primarily to addressing this question by arguing that it is unlikely that emotions are best understood as basic and, on the basis of this uncertainty, cautions against pinning the prospects of a Sentimentalist analysis of evaluative or

---

moral concepts on the outcome of the debate over basic emotions. It is better to assume that emotions are not basic.

There was once thought to be a significant amount of evidence that indicated that some emotions are basic. I argue that this evidence should be much more carefully evaluated. Much of this evidence has come in the form of 'judgment' studies on facial expressions. There is reason to think that much of the evidence that shows wide cross-cultural recognition of emotional facial expressions results from the methodology of the judgment studies themselves. When these effects are accounted for the evidence from judgment studies is compatible with the view that emotions are not basic.

Furthermore there has been little evidence supporting the claim that emotions can be identified via distinct sets of physiological profiles or distinct patterns of activation in the brain that correspond to basic emotions. Though of course all empirical results are preliminary and views on whether emotions are indeed truly basic must await a full accounting of the empirical results, I argue that evidence currently available points toward emotions not being basic and that philosophers would do well to avoid basing key aspects of their Sentimentalist theories on the question of whether basic emotions exist.

1.3 Sentimentalism

Emotions are best understood on the Psychological Constructivist's picture. On this view emotions are not basic though perhaps some of their components may be.\(^5\) I

---

\(^5\) James Russell and Lisa Feldman-Barrett, for example, argue that 'core affect,' the source of an emotions feel or charge, is likely physiologically basic and universally shared. Core affect however is only one component among many in an emotional experience. See their “Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant” in the
turn my attention toward an examination of how philosophers have made use of emotions when constructing moral theories. Here I begin by defining philosophical Sentimentalism more precisely. If Sentimentalism is to be seen as a distinct view on the relationship between our emotions and values (or moral judgments) then this relationship must not be compatible with other theories of value or morality.

Following Kant, the deontologists, for example, believe that morality can be understood as a set of a priori truths about how rational agents must view their own ends simply in virtue of their capacity for rationality. Moral duties are those duties that all rational agents could rationally will to hold one another accountable for within a Kingdom of Ends. Emotions, though they may be present for us when we make moral judgments, are best seen as distractions by the deontologists. Acting from our emotions can lead us to act in ways other than how we ought to act. The Sentimentalist conception of morality therefore should not be able to be incorporated within a deontological framework if the emotions are going to play a central role in explaining values.

In the third chapter I argue that emotions should be seen as being able to play one (or more) of the following roles within a moral system. Emotions could play a motivational role. They could give us that special 'oomph' that gets us to act rightly (or wrongly). If we are torn between doing the right thing or doing the wrong thing then feelings like guilt could motivate us to do the right thing. Emotions could also play an epistemological role. On that role the fact that we experience one emotion versus

another, for example, could be cluing us in on our moral duties or to violations of those duties. Feeling guilty because you didn't share the last cookie, for example, would be an indication that you have failed to be beneficent. Emotions then can not only help us *do* the right thing they can help us to *see* what the right thing to do is. These two roles for emotion are compatible with many other moral systems, including some forms of deontology. Because of this I argue, ultimately, that Sentimentalists must see emotions as playing a distinctly *constitutive* role. If emotions play a constitutive role in morality then there is something about the emotion itself that makes, as opposed to merely corresponds with or motivates, something moral. I focus my attention on this constitutive relation and its related problems in the final three chapters of the dissertation.

There are two different ways in which Sentimentalists have approached the issue of constitution: first- and second-order. On the first-order understanding of Sentimentalism what matters are the emotions someone, perhaps a suitably idealized someone, *actually* feels. I take Hume to be a classic example of the idealized first-order Sentimentalist. According to the second-order Sentimentalist it is not our actual emotional responses that make something moral but instead those emotional responses that are, or would be, 'appropriate' or 'rational' regardless of whether anyone actually feels them. second-order Sentimentalism has several important advantages over the first-order variety and I ultimately focus most of my attention on the second-order iteration of the view.
1.4 Two Problems for Second-Order Sentimentalism

Although both first- and second-order forms of Sentimentalism have their own virtues and vices I argue that the second-order variety holds the most promise. In the fourth chapter of the project I raise two worries for second-order Sentimentalists that threaten to undermine the view. Since second-order Sentimentalists hold that only appropriate emotions play the constitutive role for at least some of our values. I focus first on examining how second-order Sentimentalists have understood and made use of emotion within their theories. In particular it is important to concentrate on one especially promising iteration of the view: Rational Sentimentalism. Rational Sentimentalism is a second-order Sentimentalism where emotions serve two important roles. One is the familiar constitutive role. The uniqueness of Rational Sentimentalism can be found in the analysis they provide of an emotion's being 'appropriate.' Appropriateness is, they argue, to be understood in uniquely emotional terms. That is, although emotions are subject to many different kinds of assessments (that they are rational, that they are rude, that they are in our interest to express or not express, etc) the only assessment that makes an emotion distinctly appropriate is an assessment of whether an emotion is 'fitting.'

To figure out whether an emotion is fitting, the Rational Sentimentalists argue, we need to turn to the emotion itself for guidance. Each emotion, on this view, has a distinct 'locus' of concern. Earlier we understood this as the distinct kind of situation that seems to suit each emotion. Happiness, I said above, seemed appropriate in cases
where good things happen to you. An emotion's locus of concern is the evolutionary product of our shared emotional development. Each emotion has evolved to be sensitive to very specific kinds of inputs as a result of universal problems that each emotion has evolved to respond to. Fear, for example, may have evolved to respond to specific kinds of threats (those which we can do something about). According to this brand of Sentimentalism an emotion is appropriate only when we are faced with a situation that falls within the scope of the locus of concern for that emotion. If this is not a threatening situation where my agency can play some important role then fear, despite anything else it could be, is not appropriate.

Fitting emotions are then used to play an interesting kind of 'regulative' role by the Rational Sentimentalists. Knowing that an emotion is fitting will tend to keep the emotion going while knowing that an emotion is not fitting, that it is inappropriate, will tend to extinguish the emotion. This helps to make sense of, on their view, stubborn emotions that do not go away even though it would be in our best interest for them to do so. Those stubborn emotions may turn out to have been so hard to get rid of because they were, in fact, appropriate. I focus much of my attention on the Rational Sentimentalist's understanding of emotions and I raise two problems for the view on my way to amending this view into a more Sensible Sentimentalism.

Rational Sentimentalism relies on an understanding of emotions where emotions are 'basic.' A basic emotion, as noted earlier, is an emotion that is nearly universally shared at least with other human beings. The universality of the emotion is
explained in terms of a phylogenetic story. Humans share emotions in the same way we share bipedalism and not in the same way in which American pop music might be universally shared. We share bipedal movement as a result of the evolutionary advantages that accrued to our proto-bipedal ancestors. Pop music on the other hand is shared because of the particular cultural and political history of the United States relative to the rest of the world. I think that a commitment to basic emotion is risky and ultimately unnecessary.

But the Rational Sentimentalists place basic emotions on a high pedestal. We are stuck with our emotions because of the advantages they provided for our ancestors. Since they play such an important role it is best to take them seriously. Further, if emotions are not only basic but identifiable in terms of specific patterns of physiological markers then basic emotions can also help to explain important features of our evaluative discourse. Since basic emotions are not as well supported as once thought, I argue that it would be advantageous to unhitch Sensible Sentimentalism from these kinds of commitments.

Dropping the commitment to basic emotions leaves an argumentative gap for the Rational Sentimentalists. This problem can be overcome though it comes with costs. I argue that adopting the Psychological Constructivist view of emotion need not diminish the importance of an emotion nor does it mean that Sentimentalists are left without an account of evaluative discourse. Adopting such a view will have an impact on how we understand the terms of our evaluative discourse. If emotions are
constructed out of the various components described earlier (beliefs, desires, feelings, actions, and so on) and further if none of these components are necessary for any particular emotional experience then it seems as if any moral concepts constituted by those emotions will also fail to be universal. Our moral demands would be limited, on this kind of view, to those who share enough of our emotional concepts to understand the resulting moral demand. This places a kind of strict limit on whom we can engage in genuine moral disagreements with. Although this feature of the Psychological Constructivist approach to emotion can be seen as a cost, I believe any Sensible Sentimentalism must take this kind of form.

After raising this issue for the Rational Sentimentalist I go on to raise what I consider to be a more fundamental problem. Sentimentalism is defined by the fact that it places emotions in the constitutive role with at least some of our values. The constitutive role needs motivation however. Why should we be inclined to accept that any emotions are constitutive of values if emotion-independent understandings of those values can make sense of much of the same phenomena? The final two chapters are concerned predominantly with analyzing both the constitutive role and the arguments that have been offered in its favor.

1.5 Emotional Constitution

I end the project by casting my net far and wide in the hopes of snagging useful arguments for the constitutive role. Although I come up dry I find the failure instructive and argue that there are resources that help point the way to providing an
argument for the constitutive role. I canvass two different families of literature where we should expect emotional constitution to play a role. The first is in the more straightforward metaethical discussion of Sentimentalism I have already been discussing. Here I come up short.

For most Sentimentalists the constitutive thesis functions as a starting point from which theories are built. Too often emotional constitution is assumed to be the case instead of argued for. While this makes some dialectal sense, Sentimentalism can seem attractive for its ability to make sense of many different aspects of moral discourse, it does make it difficult to assess the independent plausibility of the constitutive claim itself. The Rational Sentimentalists come close to arguing in support of emotional constitution. In their case the constitutive claim results from the fact that emotions are understood as non-cognitive and because of this, in their eyes, the emotions are also uncodifiable. If emotions are uncodifiable then it is unlikely that the contribution that emotions make to our moral concepts can be specified independently of the emotions themselves and hence the role for emotion is constitutive.

This, for reasons we'll see in the penultimate chapter, is a bad argument. Other Sentimentalists have appealed to aberrant agents to support the constitutive claim. These argument also, I argue, do not get us to the constitutive claim. Lastly there is a distinctive history in the philosophy of value to analogize values with color properties. If values are like colors then there is a case to be made for the emotions to function as part of the meaning of evaluative properties. Although an interesting and historically
significant debate, I argue that the analogy is a bad one and that constitutiveness cannot be purchased with it. I argue that a fully constructivist theory of concept formation, one that is sensitive to the inter-subjective demands of evaluative discourse, can secure the constitutive thesis but at a cost. On a fully constructivist theory of concept formation all concepts would be constituted by our responses. Although constitutiveness is possible on such a view I argue that the ideal Sentimentalist story need not be so revisionist.

I then canvas a second potential source of support for the constitutive claim within what are sometimes called 'reactive attitude' theories of moral responsibility. On a reactive attitude theory of moral responsibility what it means to say that someone is a responsible agent is for that agent to be the appropriate target of a reactive attitude. Just what it is that reactive attitudes are and what makes them appropriate will vary depending on the theory but what is important for my purposes is that reactive attitude theories of moral responsibility are second-order Sentimentalist theories with commitments to the reactive attitudes playing a constitutive role. Here the arguments for constitutiveness fall short but are a bit more promising.

Most reactive attitude theories of responsibility turn to similar forms of argument as Sentimentalists generally when it comes to motivating the constitutive thesis, or so I will claim. Appeals to basic emotions and the ineliminability of our attitudes are common and founder for reasons similar to those offered against the more general Sentimentalist claims. Those theories that reject basic emotions are more
promising and appeal to cultural differences in the reactive attitudes used to enforce pro-social norms as evidence in favor of constitutiveness. These cultures, sometimes called shame-cultures, will help to ultimately motivate the constitutive claim. Although appeals to psychopathy appear once more within the reactive attitude tradition I think these appeals can be built up into a more successful argument for constitutiveness but only in combination with the claims about shame-cultures described above.

Although most reactive attitude theorists, like the more encompassing Sentimentalist theories, assume the truth of the constitutive role I argue that, at least with the concept of moral responsibility, we can begin to form an argument that motivates the plausibility of the constitutive claim. This argument takes the form of what I will call the Empirical Test for the Reactive Attitudes (ETRA). ETRA will need help from several claims to get off the ground. Psychopathic agents motivate at least the claim that the reactive attitudes are playing an epistemological role. Psychopathic agents are special because they lack just the kinds of affective capacities that appear necessary for a proper understanding of responsible agency and moral understanding more generally. Indeed their strange behavior bears out this lack of understanding. Furthermore we can make comparisons between cultures that enforce their norms in terms of reactive attitudes (here understood in terms of resentment, indignation, and guilt) and cultures that enforce norms in terms of other attitudes (like shame). Sociological information of this kind should be useful for the reactive attitude theorist.
At the close of the final chapter I will argue that, when combined, these two sources of evidence help fuel ETRA and motivate the thought that the constitutive claim may hold at least for the subset of attitudes that constitute conception of responsible agency. ETRA, as its name implies, is an empirical test and so can also be used by the detractors of constitutiveness. If there are not any good test cases to be found then ETRA itself will suggest that constitutiveness may not hold for a particular evaluative property.

The focus of ETRA is that if in both the case of aberrant agents and also in the case of agents in 'non-reactive' cultures we find distinct practices and conceptions of the kind of agency thought 'appropriate' to enforce then, I suggest, it becomes more plausible to think that the relationship between emotion and responsibility is more than merely epistemological. Were it merely epistemological we should expect more convergence. This line of argument has the added benefit of being testable, something often missing within the context of moral theorizing.

**Conclusion**

Emotions can be asked to play many different roles by philosophers who hope to address contemporary problems in moral philosophy. Before we can begin the process of examining the soundness of the moral and evaluative arguments that make use of emotion we need to be clearer about how emotions are understood and the specific roles they are being asked to play. This becomes all the more important within the broad class of theories of morality that are known as 'Sentimentalist' theories. If a
Sensible Sentimentalism is to be offered then it must steer clear of the pitfalls that have snagged earlier attempts at offering theories of morality grounded on emotions. Just what these pitfalls are and how they are tied to different theories of emotion and the different roles they have been asked to served will be made clearer in later chapters.

In the proceeding chapter I will be concerned primarily with a bit of ground clearing before beginning my argument. Theories of emotion abound and continuously multiply. I will attempt to describe and analyze the major camps that these theories have tended to fall into. I end the chapter by proposing that one particular family of theories, known as Psychological Constructivism, is best suited to explain our experience with emotion.
Chapter 2

On the Very Concept of Emotion

Sentimentalists argue that emotions play an important role in morality. Before we can ask that emotions play any particular role within our theories however we must first have a better understanding not just of what emotions are but why it has proved so hard for any particular theory of emotion to win widespread favor. Emotions, to echo Ronald de Sousa’s poetic analysis: “are on the frontier of the subjective and the objective, as they are on the frontier of the mental and the physiological, of the active and the passive, of the instinctual and the intentional. None of these frontiers is sharp: the problem is to map them.”

1 de Sousa, I think, provides us with an accurate diagnosis of the current state of the field in the philosophy of emotion. It is because emotions, whatever they are, have such distinct but heterogeneous properties that analyses of emotion are immediately contentious. Over the last thirty years many philosophers, psychologists, and neuroscientists have attempted to analyze or operationalize our folk conception of “emotion” with varying success. As is often the case, there are nearly as many conceptions of emotions as there are researchers. de Sousa successfully analyzes the problem inherent in developing an account of the emotions and he can also help point us toward a solution. Although theories of emotion have raised awareness and significance of the many different facets that go into an emotional experience, conceptual confusion remains. In what follows I canvas, in broadly historical terms, some of the more popular theories of emotion before going

on to propose that emotional experience is best explained by views like de Sousa's that acknowledge the deep way in which culture impacts the meaning of affective experience. This in turn is important because of how our understanding of affective experience explains how we acquire our concepts of the various particular emotions but also of Emotion more generally. Emotions, on this kind of view, are built up from a combination of the biologically basic and the socially constructed.

Emotions, as I suggested in the opening chapter, can be asked to play several roles within a Sentimentalist ethics. Not only are emotions described as ways of accessing moral facts, what I call the epistemological role, they are also seen as highly powerful incentives towards acting, what I call the motivational role; I have also suggested, and will argue in a later chapter, that the distinctive feature of Sentimentalism is that emotional responses are seen as constitutive of moral facts, what I called the constitutive role in the introduction. This makes it all the more important for any aspiring Sentimentalist to get more specific about what kind of things emotions are such that one class of things (emotions) can play all of these roles. We must also be convinced that emotions do play these roles.

The problem is with developing a coherent account of emotions that adequately respects the various aspects of emotional experiences that makes them so vexing to philosophers, psychologists, and neuroscientists while at the same time making sure not to be too inclusive or exclusive. Some states like moods seem importantly different from paradigmatic emotional experiences for example.
If the problem in developing an account of emotions is in the successful mapping of the concept then we stand to learn a great deal from past conceptions and their potential shortcomings. This is where I turn my attention next. First, I will try to develop a lay of the land, both contemporary and historical, of influential theories of emotion. I do not here attempt an exhaustive analysis of the history of philosophical theories of the emotions. My aims are much more limited. I want to give an overview of the basic motivations that have led to the creation of many of the more established and influential theories currently in fashion in the philosophy of emotion. Most camps have reductionist and non-reductionist versions of their theories and some are committed to the existence of a set of universally shared “basic” emotions while others reject this conclusion, at least as a priori. A commitment to one or the other of these assumptions carries theoretical costs that I intend to make clearer in the following chapter.

2.1 Where is the Emotion in an Emotional Experience?

The question heading this section underscores the level of complexity that an adequate answer to the question would need to incorporate. When trying to give any account of what emotions are, the most obvious place to begin is with our everyday emotional experiences themselves. But how should we understand these experiences?

Historically, emotions have often been compared with colors. One natural place to begin an investigation of our color concepts is with our color experience. Is there anything (physical / phenomenological / perceptual) that unites all our

---

2 I turn to the question of Basic Emotions in Chapter three.
experiences of redness? Naively we might say that it seems that all red things are disposed to cause red color experiences in us. The experiences of redness are united by the fact that they are all connected by a common cause: something's actually being red. Although an analysis of our color terms should not end at this level, any analysis of the objects of experience should take the phenomenology of those experiences seriously.

On a similar note we can ask ourselves whether there is anything that unites our emotional experiences (of disgust, anger, outrage, fear, etc). Any account of what emotions are must be able to account for emotions as we experience them. Using our experience as a starting point, while a natural way to begin an argument, brings with it its own set of problems. It may be the case that our experience does not give us accurate insight into the ‘real’ nature of emotions. If this is true then a theory of the true nature of emotions is revisionist when it comes to emotional experience. An account of emotions with revisionist aims must explain what emotions really are and how we come to mistakenly believe that they involve aspects of experience that are extraneous or in some way not actually a part of the emotion itself. While there may not be a consistent and agreed upon set of emotions, our shared sense of emotional experiences in general can serve to ground the subject matter.

Because of these difficulties, more traditional accounts of emotions have sought to explain only the nature of our emotional experiences. Neither project is necessarily limited to giving a purely descriptive answer of course. Emotions are
normally experienced as sensitivities to normative (albeit not necessarily moral) properties and this too must be explained or explained away in our theories. I’ll begin with the most basic kinds of emotional theories. These theories identify some particular aspect of an emotional experience as the emotion and explain all other aspects of emotional episodes as either antecedents of the emotion or as effects of the emotion.

The following narrative represents a typical emotional experience; I refer to this narrative in order to describe and differentiate theories of emotion.

The Interview

You sit in a small reception area surrounded by three other prospective employees. The room is stuffy and lacks windows. This is your first major job interview since graduating from college. Surveying the room you notice that the other prospective employees are sweating and looking around. Some tap their feet nervously; others rehearse their memorized talking points to themselves. Unbeknown to you, at this moment your brain is elevating the levels of certain neurotransmitters in your brain and also of certain hormones in your bloodstream. As each candidate is interviewed you note that they all leave smiling confidently. One even mouths the phrase: “This job is MINE,” on her way out of the interview office. Suddenly you realize you have butterflies in your stomach and your thoughts inevitably turn to the unknown: what kinds of questions will you be asked? Are your qualifications good enough? Probably not you think to yourself. How will the interviewer judge your merits? You assume the worst. You being to sweat. You realize that this interview is extremely important to your future prospects and you judge that you aren’t likely to get this job. The butterflies become more noticeable and distracting; then it dawns on you: you’re anxious, maybe even scared. This realization makes you jittery. During your interview you stumble when answering questions and blame your anxiety for your failure.

3 For example, experiencing an episode of fear carries with it (usually) a relatively simple normative demand: fight or flee!
The narrative is depressing but instructive and should contain elements familiar to anyone that has experienced this or most other emotions. I take it that the kind of anxiety being described is neither extreme nor insignificant and so represents a typical emotional episode. Because of its typicality this episode will serve as a ground for future discussion regarding both the elements of emotional experience and the concept of emotion itself. I want to separate six components of this narrative that are taken by many to be important to emotional episodes.

2.2 Perception of Objects

Many elements of The Interview involve perception.\(^4\) For example, there were perceptions of the office, of other interviewees, of the emotional states of the other interviewees, and perceptions of one’s own states (of various kinds including one’s own emotional states). Perception is an important element in nearly every theory of emotion. One important difference among theories is in the level of cognitive processing and access to consciousness that perception entails. It matters of course what is being perceived and what the content of perception is taken to be but specifying these parameters requires that we make commitments about the nature of emotions. At this level of abstraction it is impossible to say anything useful regarding perceptual theories of emotion. I return to perception in a moment but first I want to consider another common feature of emotional narratives found in The Interview.

2.3 Physiological Changes

\(^4\) We are, by and large, visual creatures. I do not, however, intend to limit my sense of ‘perception’ to merely visual perception (though visual perceptions are perhaps most salient). We may perceive (i.e. hear) the beating of our own heart and feel it beating faster within our chest (for example) without seeing any of those things.
Emotional experiences are often described in everyday language and make use of purely visceral descriptions to pick out and differentiate between specific emotions. We flush with excitement. Our blood runs cold in terror. We become hot tempered. Our heart drops at the announcement of bad news. In the narrative above I included some aspects of this kind of description. For example there was talk of emotional butterflies, sweating, and jitters. One, mistaken, way of characterizing emotions is to identify the emotion itself with whatever patterns of physiological changes correspond to our visceral descriptions. This kind of view would equate our talk of butterflies in our stomachs, for example, with whatever is going on in our bodies during such an episode. This characterization of emotions identifies them with various physiological changes and is one member of a family of emotional ‘physiological’ theories. The “butterflies” in The Interview would, on this account, be explained as a metaphorical expression of the changes in neurotransmitter ratios in the brain, hormone levels in the blood, blood-pressure changes, heart-rate changes, skin conductance or arousal spikes, core-temperature changes and so on that happen at the time butterflies in the stomach are experienced and that are thought to be responsible for, and experienced as, butterflies in the stomach.

In order to be able to identify each emotion type most proponents of this kind of view look for unique physiological profiles for each emotion. Such a step is necessary because if each emotion could have different physiological profiles on
different occasions then the only way to confirm that one and the same emotion is
being measured in both cases despite different physiological profiles is by appealing to
some further facts. For example we may need to turn to reports from individuals
experiencing the two states and their judgments that the two states represent the same
emotion. What is so damning about these reports is that these further facts would not
be facts about the *physiology* of emotion. Appeals to further facts would thereby
undermine the basic commitment of a pure physiological theory. It is unlikely that
distinctive physiological profiles exist for the states we normally consider emotions.
Anxiety, for example (and depression, anger, disgust, envy, jealousy, etc) has not been
shown to have a distinctive physiological profile despite early claims to the contrary.⁶
This view faces more than merely empirical problems however. It also seems that this
view would allow one to be in several different emotional states, perhaps even in rapid
succession, without us ever being aware of this fact about ourselves. Recall that this
kind of feeling theory does not require an agent to be conscious of the actual
physiological changes that emotions are taken to instantiate (we need not be conscious
of an increase of acetylcholine or any other neurotransmitter). They are also
independent from the phenomenological perception that physiological changes would
induce in a normal human being. While not logically impossible this makes this theory
revisionary in a way that should at least give us pause and tempt us to look for other

physiological emotional markers.
options. Some have thought that this worry can be addressed by combining two aspects of emotional experience: perception and physiology.

2.4 Perception of Physiological Change

Physiological changes by themselves may have seemed attractive as an account of emotions for several reasons. First, as I said earlier, physiological changes certainly seem like they are involved in our everyday emotional experiences. Second, most animals are capable of experiencing emotions on a physiological account because much of our physiology is shared with other animals. This kind of theory however also ran into empirical and conceptual problems. One attempt at resolving the problems with pure physiological theories combines the physiological component with a perceptual aspect. While technically a hybrid account because it combines the pure physiological view with a perceptual one, I treat this as a more sophisticated physiological theory because of the importance of physiology in explaining and identifying particular emotions. This modified perceptual account was first popularized by William James. James’ theory has been the progenitor of many responses both friendly and antagonistic. In The Interview there was a point when ‘suddenly’ we became aware of the butterflies and were able to locate them in a physical location: our stomach. If we carry over our earlier assumption that

7 Indeed this attraction explains why so much of the research on emotion involves doing all manner of things to rats. The assumption being that the physiological changes associated with emotion are evolutionary and should be found in most mammals (at least). Whether or not emotions are reducible to those physiological changes is another issue.

“butterflies” in the stomach are important aspects of an emotional experience and that they are grounded on changes in physiology, we can then add conscious awareness of these changes as a necessary condition to give us James’ view. James argues that physiological changes precede emotional awareness and that awareness of these changes in our physiological state constitutes what it means for something to be an emotion. Once I have noticed that I have butterflies in my stomach then, and only then, am I actually experiencing anxiety. Anxiety just is my becoming aware of the butterflies. Fear just is the particular feeling derived from experiencing an increase of adrenaline (or whatever else is required) into the bloodstream. This solves one problem raised by the pure physiological view but leaves untouched the other, empirical, worry about distinct physiological profiles for each emotion.

Finding distinct physiological profiles for at least some class of emotions (the “basic” or “atomic” ones) is necessary in order for a view like this to account for the richness of human emotional experience. This is, I am willing to bet, unlikely to happen. For the same reasons that the pure physiological theory have trouble differentiating individual emotions with an appeal to distinct physiological profiles, James’ modified perceptual feeling theory will run into the same set of problems. It does however have one crucial advantage.

With further modification, a theory like this could drop its commitment to emotion specific physiology entirely and instead place more weight on the perceptual/phenomenological aspect of the emotional experience. Any physiological
state that makes us feel euphoric would do just fine so long as we feel euphoric. A view like this certainly seems to have an advantage over the earlier one but it too is apparently in danger of running afoul of actual emotional experience. We cannot feel euphoric at a sudden drop in blood pressure for example. The link between physiology and emotion is not completely ad hoc and this should be a telling feature of the nature of emotions in general. Another objection to feeling theories is that physiological changes are neither necessary nor sufficient for us to have emotional experiences.\(^9\) Antonio Damasio, for example, has argued that physiological changes can be simulated by the brain itself using what he calls 'as-if' loops that have the same effects as the physiological changes, and are experienced as being similar, without actually requiring a change in hormone levels, or blood pressure, etc. Emotions are originally perceived as physiological, according to Damasio, but the brain can come to simulate these physiological states by accessing the relevant brain areas that would be activated by physiological changes even when those changes are not present. This is an as-if loop. It is 'as if' we are experiencing the physiological change. Damasio in fact argues that most emotions occur via these ‘as-if” loops in the brain and not via bodily changes. If true then Damasio's observations suggest that what we are perceiving when we have an emotional experience is, much of the time, not a bodily change at all. It might be better, some have thought, to focus on another aspect from The Interview, an aspect that does not make unattractive empirical commitments to distinctive

physiological profiles: emotional thoughts.

2.5 Cognitivism about Emotion

Thoughts are prevalent throughout an emotional episode: the boss hates me, I am worthless, that person over there is very attractive, I have just achieved one of my life’s goals, etc. In The Interview our thoughts were primarily focused on various features of the situation: the room, other interviewees and their reactions (emotional and otherwise), assessments of the interview, physiological changes, self doubts, etc. A purely cognitive account of emotion claims that beliefs are the only necessary component of an emotional episode. Mixed cognitive accounts make the much weaker claim that beliefs are necessary for emotional states but allow for other states to be necessary as well. I will consider these mixed cognitive theories when I discuss hybrid accounts of emotions. What all cognitive accounts share is the claim that beliefs are necessary and that they have pride of place in a theory of emotion. If I did not believe that I might fail to land a job, a prospect that I also judge to be very bad for myself, then I simply have not yet experienced anxiety. More generally (and accurately), if I did not construe the interview as an important goal and also construe my chances of attaining that goal to be slim and, possibly, that this is because of a personal failing, then I am not genuinely experiencing anxiety. On a pure cognitive account emotions are differentiated only by the beliefs or judgments that an emotion is identified with. This requires that emotions be identified with specific patterns of judgments or thoughts.\footnote{Were we to construe the loss of the job as being due to interviewer bias instead of a personal failing,}
constitutive thoughts regarding a major task or obstacle and the possibility of personal failure. Fear might be constituted by the beliefs that something is both dangerous and an imminent threat to the safety of my loved ones or to myself. Envy, on such a view, might be identified with the belief that a rival has received a benefit that I wanted for myself. Pure cognitive theories of emotion thus make things like beliefs necessary and sufficient for emotional experience. Beliefs are one kind of mental state that can render a theory cognitive but other options are available.

Cognitive theories of emotion require that agents have propositional attitudes directed at an object. A belief that the bear poses a threat to me for example expresses a proposition (that the bear is a threat) and a kind of attitude that we direct towards it (belief). Cognitive theorists disagree about the kinds of propositional attitudes that are necessary in order to label a state an emotional state. For example, Robert Solomon argues that emotions represent specific kinds of judgments that are not reducible to beliefs and are not necessarily conscious:

Judgments, unlike thoughts, are geared to perception and may apply directly to the situation we are in, but we can also make all sorts of judgments in the utter absence of any object of perception…judgment seems to me to have the range and flexibility to apply to everything from animal and infant emotions to the most sophisticated and complex adult human emotions such as jealousy, resentment, and moral indignation.¹¹

More specifically:

An emotion is rather a complex of judgments and, sometimes, quite

for example, we might be in a very different emotional state (outrage or resentment say).

sophisticated judgments, such as judgments of responsibility (in shame, anger, and embarrassment) or judgments of comparative status (as in contempt and resentment). Emotions as judgments are not necessarily (or usually) conscious or deliberative or even articulate but we certainly can articulate, attend to, and deliberate regarding our emotions and emotion-judgments, and we do so whenever we think our way into an emotion, ‘work ourselves up’ to anger, or jealousy, or love.¹²

Solomon, and cognitivists like him, prefer to talk of cognition in a broad sense that includes all mental states, conscious or not, whose intentional content focuses on important aspects of the primary concern of an emotional episode (its object). Another similar approach borrows the idea of multiple judgments or propositional attitudes in order to explain our emotions. This approach is most often made by psychologists who argues that emotions are composed of judgments along several dimensions. This family of views are called dimensional appraisal theories. These dimensional appraisal theorists disagree on the number of dimensions or judgments that must be made to differentiate one emotion from another but their analyses are similar in the sense that what distinguishes emotions from one another are the different sets of appraisals associated with each. The simplest dimensional appraisal theories, for example, argue that emotions result from an agent’s assessment of a situation in terms of her needs, goals, desires, and beliefs.¹³

A common objection to cognitive approaches in general is that they violate our everyday experience with emotion. If fear, for example, is a judgment or belief or appraisal that an object or state of affairs is dangerous, then it appears that one often

¹² Ibid (11)
can and does believe that something is dangerous without necessarily being afraid of it. For example, we may acknowledge that cigarettes and risky sexual practices are dangerous and yet we might largely fail to find these things fearsome. In fact, as we will see in chapter three, some press the objection even further to argue that cigarettes and risky sexual practices are simply not appropriately viewable as fearsome. So long as that statement makes sense to us then the belief or judgment or appraisal that something is dangerous can be disassociated from the experience of fearing the objects of those propositional attitudes. Beyond this, it also seems that we often find ourselves being afraid of things we know are not dangerous. Many people have a genuine fear of riding roller coasters or flying on airplanes despite the fact that they also acknowledge that these activities are statistically safer than traveling in cars (which fail to elicit fear from these same subjects). Cognitivists do have responses to these worries. These involve attributing further beliefs to agents in order to explain away the apparent contradictions. Of course a simpler response to these worries, and one that is intuitively more plausible, is that emotions are not simply reducible to propositional attitudes but instead that they must make reference to other mental states: conative states. Another way out of this problem is to eschew appealing to propositional attitudes at all. To the degree that these “noncognitive” theories or hybrid cognitive-conative account of emotional experience can make sense of our actual experiences without attributing false beliefs to us then that is a reason to favor such alternatives. I

---

do not of course take myself to have offered an irrefutable argument against pure cognitivist theories of emotion. Rather, I take myself to have motivated the search for a theory that better accounts for the set of emotional experiences available to us. A purely cognitive account makes our beliefs not only dominant but also the only elements of consequence in a theory of emotion. This fails to do justice to the full range of emotional experience. I want to turn my attention to another attempt to identify an emotion with a single component of emotional experience before considering different hybrid accounts.

2.6 Behavioral Changes

Behaviorists make it a point to eschew talk of mental states. Instead, they prefer to focus their theories on observable changes in patterns of behavior and how those patterns of behavior are caused or modified by events or other persons. It is clear that emotions generally incite us to do things. Behaviorists therefore focus on identifying the patterns of behavior generally associated with different emotions. The Interview included several changes in behavioral dispositions both within our (fictional) selves and in others. Uncontrollable tapping of one’s feet, unusual stuttering during the interview, pacing, and so on are recurring patterns of behavioral changes that are associated with nervousness or anxiety. The key insight of the behaviorists is that emotions might be identified and defined in terms of stereotyped patterns of behavior that each emotion elicits.

Historically, Darwin is perhaps the first and most celebrated behaviorist. He
noted that individuals made specific facial expressions during emotional episodes and that these expressions were often accompanied by various common patterns of behavioral responses (a frightened face followed by fleeing for example). More sophisticated behaviorist approaches tend to include specific types of cognitive changes within an analysis of an emotion. For example, sadness might be thought to normally lead people to make more accurate judgments of their own abilities, a phenomena known as depressive realism, while happiness leads to judgments made on the basis of less careful heuristics. Behaviorism has been popular with both psychologist and philosophers.

Emotions on the behaviorist's view are defined as recurring patterns of behavioral dispositions (on a significantly wide notion of what counts as ‘behavior’). There are very few contemporary behaviorists. This may be because emotions are strongly tied to the 1st person perspective. We know emotions best from our own experiences with them and any theory of emotion that neglects this distinctly phenomenological aspect of emotions seems incomplete. When behaviorists begin including changes in perception and cognition within their accounts then the theory becomes more satisfying but these inclusions come with a cost. More inclusive views like these begin to tread into the dangerous territory of counting unobservable mental

---


content as behavior, that is, they tend to betray their own behaviorism. On the other hand hybrid theories of emotion that eschew a strict commitment to pure behaviorism have incorporated behaviorist elements as important components of our conception of emotion. Because these accounts avoid the problems that plague behaviorist theories and because hybrid theories of this kind are by far the most popular approach within the philosophy of emotion, we can appreciate the legacy that behaviorism has had.

Thus far I have discussed theories of emotions that selectively favor one aspect of emotional experience over others and call this favored aspect the *real* emotion. Other factors that precede or proceed the chosen aspect are taken to be either causes or effects of the emotion but are not seen as being a part of what really matters when we talk about emotions and emotional experiences. As might be expected with this kind of approach, a common complaint is that each of these views leaves something out. The perceptual physiological theories neglects some kinds of emotions that seem divorced from (relevant) physiological changes. The cognitivist makes the opposite mistake and strays too far from the felt experience of emotions. What separates ‘hot’ emotions from ‘cold’ judgments, the argument goes, is the affective quality of emotional experience itself. Behaviorists, on the other hand, neglect the distinctly personal elements of emotional experience by focusing only on behavioral dispositions. For this reason most theories of the emotions are what some have called hybrids of some or all

---

18 This is not a damning objection. It would be easy for the behaviorist to reinterpret these apparently internal states in terms of purely external behavior. For example depressive realism should be understood, on this kind of reinterpretation, not as a kind of experience but purely by reference to the kinds of reports depressed persons make. While this saves behaviorism it remains open to the original charge: that leaving out the feeling component of an emotion is revisionary.
of the elements above. Hybrid theories are certainly in a better position to account for our emotional experience than any of the singular views but, as we shall see, they have problems of their own. The most common hybrid in the field marries beliefs or judgments with feelings, desires or behavioral dispositions. Ultimately I will make a case in favor of a particular kind of hybrid account, the Psychological Constructivist theory, on the grounds that it best explains emotional experiences. I turn my attention now to an examination of hybrid theories including the currently received view of emotion.

### 2.7 Hybridization

The preceding theories share two important features. All of the theories took one element from emotional experience and crafted a conception of Emotion centered around that one element (to the exclusion of other elements). The second feature being that each of these theories was, not surprisingly, unable to fully account for the entire range of emotional experience. Some of these problems were empirical, for example the lack of emotion specific physiology ruled out physiological theories like James’ account. Other problems were theoretical and made some theories less attractive than others. For example the omission of feeling states from behaviorist and cognitivist theories made them more revisionist than they may appear.

Hybrid theories attempt to take the best components and arguments of all these theories in order to patch up problems and create a more complete view. Hybrid theories come in many forms. For my purposes it will be enough to limit attention to
two general hybridization strategies that serve to distinguish classes of theories from one another. I call one of these strategies the all-components-necessary strategy (ACN). An ACN approach gathers together a collection of components (action tendencies, beliefs, desires, physiological changes, and so on) and defines an emotional state as that state in which all chosen components are present and structured appropriately. For example, a very common kind of hybrid joins cognitive theories of emotion with physiological and behavioral theories. This hybrid theory is the received theory of emotion in the field today. While many of theorists within this camp disagree on the stress that should be laid on each of the components, the order in which each occurs, and the exact nature of the thoughts, feelings, and behaviors involved, they do all share an important feature that unites them under as ACN theories. All the components of an emotion need to be present and structured in a particular way in order for a state to be an instance of a genuine emotion. This strategy has obvious advantages that have attracted so many philosophers and psychologists to it. For example, if the problem with the purely cognitive accounts of emotion was that they neglect to include phenomenological aspects of emotional experience then an easy way out of this problem is to argue that emotions are beliefs coupled with

---

feelings of the right kind. Justin Oakley offers a concise example of this kind of hybrid theory. His particular view requires two propositional attitudes: beliefs and desires along with affective experience:

On my view, fear would be analyzed as a cognition of imminent harm to someone and a desire to avoid this harm, along with certain appropriate affects, such as being alarmed or feeling ‘hyped up’. Anger I would take as equivalent to being pained by the cognition that we (or others we care about) have been injured or wronged, and having a desire to retaliate against the offender.20

All theories of this nature must go on to structure the relation between their individual components. This can sometimes prove to be a difficult challenge. Again Oakley proves instructive:

So far I have been describing emotions as combination of cognitions, desires, and affects. However, putting the view in this way may give the impression that I am proposing that when all the appropriate elements of a certain emotion-type simply coexist in us, then we actually have that emotion. But this is false. For if we have each of components of a particular emotion but these components have radically different causes, then we do not in this case have that emotion.21

Suppose, for example, that a neurological imbalance has led to a release of adrenaline in your bloodstream. Although causally unrelated, this event happens at the very same time as you are watching George Romero’s classic 1979 version of *Dawn of the Dead*.22 Furthermore, at this moment you drop the drink you have been holding and as a result of a startle response you jump away as it falls. You sit there, wide-eyed, heart pounding, as zombies fill your screen. A case like this would seem to have all the

---

20 Oakley 1992 (15)
21 Justin Oakley 1992. (15-16)
elements of an emotional experience: physiological changes, behavioral responses, and judgments about how terrifying the zombie apocalypse would be; but, Oakley would argue, this case would not constitute an instance of genuine fear because the elements are not properly structured: they are not all caused by one object that you fear.\footnote{The appropriate elements in this case being: thoughts about flesh-eating zombies, physiological changes, and behavioral changes that correspond with ‘flight’ behavior.}

Oakley helps us to see a further requirement that hybrids must satisfy. They must give an account of the relations that hold between the various components of an emotional episode in order for it to be a genuine emotion (as opposed to a false \textit{Dawn of the Dead} emotion). Jesse Prinz has poetically labeled this the ‘Problem of Plenty.’ The problem with the Problem of Plenty is that philosophers must show how each of the component parts of an emotional episode is structured into a coherent whole: the emotion.\footnote{Jesse Prinz \textit{Gut Reactions: A Perceptual Theory of Emotion} Oxford University Press, 2004} They must also explain why that particular structure should be taken as the canonical structure for each emotion. In other words, given that these philosophers believe that all the components mentioned are necessary in order for something to count as a genuine emotion, a theorist needs to go on to show how emotional experiences are structured and where the components fit within this structure. Oakley provides one unifying response by creating a \textit{causal requirement}. If the same object or state of affairs does not cause an agent to form beliefs and desires \textit{about} that particular object and furthermore also have an affective experience \textit{because of it} then the agent is not experiencing a genuine emotional episode. Before moving on to discuss the second
hybridization strategy I would like to make a general comment about ACN theories.

Returning to Oakley’s example of an ACN theory, one can ask two related questions about any theory of this kind. The first question can be directed at the subject of an affective experience that consists of some, but not all, of the emotional components thought to be necessary. Do subjects tend to report those experiences involving only desires and an affective component as genuine emotional experiences? What about experiences involving beliefs and an affective component? Beliefs and desires? Affective components by themselves? If the answer is yes, that subjects do in fact label all of these experiences as instances of genuine emotional experience, then the question for the all-components necessary view is to explain why people should be seen as making a mistake in their judgments and indeed why it is a mistake in the first place. There is some evidence that subjects do in fact make these kinds of judgments about experiences that lack one or more components of a typical emotional episode. This makes all-components-necessary theories more revisionist than they might initially appear. There is little doubt that the ACN strategy for hybridizing theories of emotion is a popular one that has many advantages over the single component theories. These views must however still address some important hurdles. ACN views are revisionary. They would render some of our judgments about what counts as an instance of an emotional episode as false. An explanation is needed that would explain

---

25 James Russell goes to great lengths to argue that our actual use of emotion concepts (specifically anger in this case) possess the characteristics of prototypes: fuzzy boundaries, internal structure (similarity judgments can be made about the degree to which something is a good example of the concept), slower reaction time when recognizing bad exemplars of a concept. See Russell, J. A., & Fehr, B. “Fuzzy concepts in a fuzzy hierarchy: Varieties of anger.” *Journal of Personality and Social Psychology*, 67, 1994. 186-205.
why some judgments about emotions would be valid and others would not be. To avoid this burden it would help to be able to hybridize a theory of emotion without insisting that all the components are necessary.

Instead of assuming that all components are necessary for a state to count as an emotional one we might instead assume that all or some (or some group) of the components are sufficient and that none (or just less than all) of the components are necessary. Call this strategy and the family of theories that appeal to it 'some components sufficient' approaches (SCS). For example, affective experiences coupled with behavioral changes of the right sort might be taken to be sufficient for an emotional experience even though affective experiences or behavioral changes (by themselves) are not always necessary. More specifically, fear may not require thoughts about danger but only a felt adrenal response along with flight behavior (although cognitions may be present). Jealousy may require beliefs and desires about one’s relationships with others but may not require that one have any kind of ‘felt’ affective experience, at least not during every occurrence of jealousy. One apparent consequence of a theory like this is that emotions, as a class, may not have any unifying features. Although I do not wish to weigh in on this claim, it is open to the SCS theorist to deny that emotions form a natural kind. Another consequence is that emotional concepts may need to be developed in a piecemeal fashion where the significant features of each particular emotion must be identified without the aid of an

---

26 This becomes more puzzling when I examine the way judgment studies have been used by many ACN supporters to validate claims about basic emotions.
overarching higher level category that unites them.

This strategy has not been as popular amongst philosophers or psychologists and it seems like Prinz’s Problem of Plenty is especially devastating against these SCS strategies. Emotions, on this view, are difficult to conceptualize. Because these views do not hold that each type of emotion has a set of necessary conditions that define it, the larger category of Emotion becomes difficult to analyze. Explaining why we group these disparate kinds of experiences together and call them all emotions requires a more lengthy and articulate defense. Given the problems faced by the singular theories of emotion along with the revisionist and empirical commitments of the ACN theories, I believe that some version of an SCS theory is best suited to conduct de Sousa’s ‘mapping’ project and I believe de Sousa himself can help us to sketch what such a theory would look like.28

My hypothesis is this: We are made familiar with the vocabulary of emotion by association with paradigm scenarios. These are drawn first from our daily life as small children and later reinforced by the stories, art, and culture to which we are exposed. Later still, in literate cultures, they are supplemented and refined by literature. Paradigm scenarios involve two aspects: first, a situation type providing the characteristic objects of the specific emotion-type…and second, a set of characteristic or “normal” responses to the situation, where normality is first a biological matter and then very quickly becomes a cultural one.29

Given the malleability of a paradigm scenario I believe this qualifies as an SCS theory of emotion and I argue that de Sousa is essentially right about emotional concepts. I also believe he is right about emotions being built out of paradigm scenarios and that

---

28 It’s unclear whether de Sousa himself would accept a hybrid theory of this sort although I believe that his general theory of emotions laid out by de Sousa (1992) is compatible with such an approach.  
29 de Sousa, 1992 (182) original emphasis.
these paradigm scenarios are, for the most part, socially constructed but have a grounding on innate human responses. While de Sousa is not more specific on how paradigm scenarios operate he is clear about the intended upshot of the view:

[The theory of paradigm scenarios] fits in with the fact that some emotions are more thought dependent than others. It depends on the paradigm scenarios to which they are related. If the paradigm scenario cannot be apprehended without complex linguistic skills, for example, we shall not expect to find in someone who lacks those skills an emotion specifically tuned to that scenario...on the other hand, prelinguistic responses such as flight or attack can subsist to define so-called primitive emotions of fear and rage. Thanks, perhaps, to their dependence on relatively separate parts of the brain, they retain their power even over individuals whose repertoire includes the most “refined” emotions.\(^\text{30}\)

de Sousa intends to fill in the gaps left by the theories considered thus far. He argues that all components of an emotional experience cannot be necessary because, as a view of this sort might indicate, emotions are not a natural kind and so do not share a set of necessary and sufficient features. Paradigm scenarios serve to define which of the many elements of emotional experiences are needed to correctly ascribe a particular emotional concept and the content of a paradigm scenario is largely the result of a social project. Furthermore, paradigm scenarios function as exemplars; they are idealized situations that show most clearly when certain kinds of emotions would be *appropriate* and to this extent they involve emotions within a normative domain in complex ways not explored by previous theories and that are only now being fully appreciated by philosophers and psychologists. de Sousa’s work has influenced many philosophers but not many have accepted his particular conception of paradigm

\(^\text{30}\) de Sousa 1992 (184)
scenarios and their constitutive role in determining emotional concepts. I believe that this is a mistake and that a theory of the emotions is strongest when it incorporates something like paradigm scenarios or, even better, adopts a prototype theory of emotion concepts. Much of de Sousa’s work is echoed by, and finds empirical support from, the research carried out by psychologist James Russell work and his influential ‘Psychological Constructivist’ theory of emotion.

Russell argues in favor of a prototype theory of emotion concepts and thus also has an SCS theory of the emotions. I believe a theory of the emotions like Russell’s is best suited to incorporate the lessons from the philosophy of the emotions with empirical data. I want to focus more specifically on Russell’s theory in order to show how a view like Russell’s helps to fill in the gap from de Sousa’s paradigm scenarios to his intended upshot.

2.8 Russell’s Psychological Constructivism

James Russell, a psychologist at Boston College, has argued against what I have called the ‘received view’ of emotions. By studying our actual use of emotional concepts, Russell comes to the conclusion that any conceptual analysis that posits necessary and sufficient conditions for emotional concepts will fail to capture our actual use of emotional concepts.31 In other words, the emotional experiences that elicit fear, for example, cannot be used to form an exhaustive set of necessary and sufficient conditions that serve to define all and only ‘fear’ states. Our emotional

---

concepts, Russell argues, are too unruly for that kind of analysis. But this forces Russell, and anyone else adopting the SCS strategy for a theory of emotion, to explain why certain experiences are conceptualized as experiences of “fear” while others are conceptualized as “anger” or “disgust”? de Sousa and Russell have suggested similar and intriguing responses to this question. They both begin by appealing to a kind of idealized concept, paradigm scenarios in de Sousa's case and prototypes or exemplars in Russell's case and then each argues that our everyday emotional experiences are understood in the context of these idealized concepts. Everyday emotional experiences are identified in relation to these concepts. If my affective experience has more components in common with the prototype for fear than it does with the prototype for anxiety then these theories predict that we will claim to experience fear. We would make these judgments even if many components of a prototype or paradigm scenario for fear are missing so long as the affective experience remains more like fear than other nearby prototypes. Russell’s view acknowledges that some aspects of emotional experience appear universal. The range of affective experiences appears to be physiologically limited. Russell also argues that emotions arise as a result of different ways in which social influences have combined the available components of an emotional experience into discrete emotions. The historical-physiological-social prototype concepts that result from this process form the emotions with which we

32 Although there are important differences between prototype theories of concepts and exemplar theories, I believe that these differences are not, in this case, as important as the similarities: both deny the all-components-necessary approach and instead rely on a model (be it prototype or exemplar) and similarity relations between that model and a sample in order to judge whether the sample falls under the model concept.
become familiar. Russell’s view makes use of these partially socially constructed prototypical emotional episodes to explain how individuals make judgments about their actual emotional experiences.

According to Russell, prototypical emotional experiences or paradigm scenarios rarely occur in everyday life. Instead everyday emotional experiences often lack many of the elements of the prototype. Our everyday emotional experiences are experienced as discrete emotions because of the similarities that our everyday experiences share with different emotional prototypes. When one prototype is a clear best fit for the experience we are undergoing then that particular emotional experience is judged to be an instance of that emotional type. When clear cases are lacking, then subjects often make explicit judgments about the relation between their current affective state and a prototype or they may remain in an ambiguous emotional state. For example, the paradigm scenario for fear might include a judgment that an object or state of affairs poses an imminent and direct threat to us or to our loved ones along with the feeling of adrenaline and flight behavior. An agent that finds herself making the right judgment and experiencing the feeling of adrenaline but whom fails to fight or to flee will likely judge herself to be afraid given the similarity between her current experience and her prototype for fear. For my purposes it is sufficient to sketch Russell’s theory in order to explain its advantages over ACN hybrids.

Several empirical findings have also led Russell to move away from the more traditional ACN received view of emotion concepts and instead to adopt an SCS
prototype account. While one could presumably hold an innate and holistic theory of emotion concepts, the dominant theory of emotions posits a set of basic emotion categories out of which all other emotions are constructed.\textsuperscript{33} As we will see in the following chapter, Russell’s research strongly undermines the empirical evidence in favor of basic emotions. This leaves the SCS theories in need of an explanation for the apparent uniformity, at least within cultures, of emotional concepts.\textsuperscript{34} Furthermore, several converging lines of research have shown that specific emotion concepts like anger and love share the features of prototype concepts: they lack clear inclusive boundaries, we make similarity judgments about members within a category—for example whether rage is a good example of the category “anger”—and we are slower to judge that less prototypical instances of an emotion fall under a specific emotion category. Subjects are slower to judge that feeling indignant counts as a type of anger than they would be to judge that rage is a type of anger, for example, even though subjects to tend to classify both indignity and rage as anger.\textsuperscript{35} Lastly, subjects’ reports of their actual emotional experiences often lack the structure posited by all-components-necessary accounts of emotions. Data suggests that the structure of our day to day emotional experiences do not have a clear cut hierarchical division. This

\textsuperscript{33} These two theories are not identical. The holist argues that it is not possible to have one innate emotion concept in isolation. The basic emotions theorist is not committed to that thesis and argues that physiological deficits can deprive subjects of some emotional concepts without a loss of emotional understanding. Furthermore, I will devote the bulk of the third chapter to the question of whether or not it makes sense to accept basic emotions.

\textsuperscript{34} See Russell's. “Is there universal recognition of emotion from facial expression?: A review of the cross-cultural studies.” \textit{Psychological Bulletin}, 115, 1994,102-141 for a critique of the basic emotions view. This discussion will occupy the bulk of the next chapter.

\textsuperscript{35} From Russell and Fehr's “Fuzzy concepts in a fuzzy hierarchy: Varieties of anger.”
places some constraints on theories aiming to do justice to our actual conceptualizations of emotional experience. Any theory of emotions that attempts to do justice to our actual emotional experiences without resorting to being revisionary must do justice to these facts about our emotional perceptions. While these facts lend some support to an SCS approach, an important question lingers. If our conceptions of our emotions do not conform to the structure imposed on them by the ACN approach, what sense can we make of the unity of our emotional responses? Russell’s answer is partially physiological and partially sociological, partially universal and partially cultural:

Prototypical emotional episodes necessarily include all of the following: core affect...; overt behavior of the right sort (flight with fear, fight with anger, etc.) in relation to the object; the experience of oneself as having a specific emotion; and, of course, all the neural, chemical, and other bodily events underlying these psychological happenings.36

On the face of it this makes Russell’s theory look like an all-components-necessary hybrid but we must be careful here. Russell does not argue that our emotional experiences necessarily must include all of these components; instead, Russell tries to makes sense of the apparent fact that our emotional prototypes sometimes seem to have these qualities and it is at the level of prototypes that all of the elements of emotional responses appear necessary. Emotional prototypes, recall, function much like de Sousa’s paradigm scenarios. They serve as reference points for what is most characteristic of a particular emotional experience. This is true even though the

---

36 From Russell & Barrett. “Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant.” *Journal of Personality and Social Psychology*, 76, 1999: 806
content of the prototype is subject to revision. It is, on this view, what we use to make sense of our affective experience. Emotions are not identical with their prototypes however and this distinction is crucial. The prototypes serve an epistemic function in that they help us to understand and make sense of our actual emotional episodes. Russell believes that “in this full blown prototypical form with all elements present, [emotions] are quite rare. (Non-prototypical cases, with one or more element missing or altered, are more common.)”\textsuperscript{37} If emotional episodes are identified via reference to the degree of similarity that they share with a prototype and the prototypes are composed of various parts, how does Russell solve the Problem of Plenty? How do the prototype components hang together? Are certain emotional prototypes more likely?

2.9 Emotional Prototypes

Russell’s later work has gone on to include at least ten components of a prototypical emotion: antecedent events, perception of affective quality in an object, change in core affect, attribution of change in core affect to object, dimensional appraisal of situation and context, action directed at the object, physiological and expressive changes, what he calls 'subjective metacognitive judgments' (sense of urgency, uncertainty, etc), an emotional label being given to the experience (emotional meta-experience), and any attempts made to regulate the emotion.\textsuperscript{38} The actual structure of any particular emotional prototype, on this view, may combine the order of these elements in various ways and many of the components may occur

\textsuperscript{37} Ibid.
simultaneously or over an extended period of time. Although I do not wish to necessarily commit myself to Russell’s particular view, his general strategy offers many advantages over those covered previously.\textsuperscript{39}

My earlier survey of theories of emotion lead us to the belief that they had all captured an element of truth. Emotions do indeed appear to include physiological changes, beliefs, appraisals, changes in behavior, self-perceptions and so on. The problem with many of the earlier theories was that these elements were believed to be \textit{necessary} for our experiences to be \textit{properly} labeled as emotional. This extra commitment about what emotional states ought to be like is revisionary and should give us pause since it accuses nearly all of us of grave errors about our own experience without explaining why we should consider ourselves in error in the first place. What Russell attempts to do is to accommodate these important insights into a theory of emotional experience and the concepts that we use to make sense of those emotional experiences. Our emotional paradigms or prototypes may appear to be expressible by a set of necessary and sufficient conditions but it is a mistake to believe that the prototype itself functions as a limiting condition on what experiences are properly counted as emotional (in fact the relation goes in the other direction). Our emotional prototypes are not guaranteed to be universal or constant—even across a single individual—and prototypes or paradigms can change as a result of experience. The

\textsuperscript{39} Lisa Feldman-Barrett, a former student of Russell, proposes various intriguing modifications to Russell’s particular theory. Her own theory is a sub-species of the SCS strategy and is a live alternative to those finding dissatisfaction with Russell’s theory. See her “Solving the Emotion Paradox: Categorization and the Experience of Emotion” \textit{Personality and Social Psychology Review}, Vol. 10, No. 1 2006: pp. 20–46.
vast differences in conceptions of emotion cross-culturally, along with the lack of evidence for specific neurological or physiological markers for basic emotions led Russell to conclude that only ‘core affect’ is a universally shared feature of all the experiences commonly labeled to be emotional and that emotional prototypes, though grounded in core affect, are largely a cultural phenomenon. At each moment every individual undergoing an emotional experience constantly updates his or her own mapping of the emotional concepts according to their own experiences and their stock of cultural prototypes. It is for this reason that Russell has labeled his theory Psychological Constructivism:

An *emotional episode* is an event that counts as a member of an emotion category, such as that labeled *fear*. A prototypical emotional episode is an event that counts as an excellent member. The proposed approach is called *psychological construction* to emphasize that the sequence of events that make up any actual emotional episode is neither fixed by biological inheritance from the human evolutionary past (as basic emotion theories have it) nor fixed by social rules or categories (as social constructionist theories have it) but is constructed anew each time to fit its specific circumstances.

Russell thus answers the Problem of Plenty in two different ways. The first response is to argue that, at the level of prototypes or paradigms, the components of an emotion hang together because each particular arrangement has historically proved itself to be a useful one for us in order to communicate about our affective lives with one another.

---

40 Core affect is simply the most basic awareness of affective states. Core affect lacks intentional objects (although when changes in core affect are registered by an agent they often search for an object to attribute that change to). Core affect is defined as a circumflex with two axes: valance and arousal. Because core affect is only a necessary (but not necessarily sufficient) component of an emotional episode, two emotions [for example fear and anger] can share the same space in core affect (they would both be high arousal negatively valenced states). Further components of the respective prototypes would then be needed in order to differentiate the two emotions.

41 Russell 2003 (151) original emphasis
While this may seem like an unsatisfying reply to many (including, as we will later see, the basic emotions theorist), Russell's empirical results and other broadly phenomenological arguments lend support to this style of theory. At the second level, the level of actual everyday emotional experience, Russell's response to the Problem of Plenty is to deny that it is a problem. Everyday emotional experiences need not have a paradigmatic structure or a paradigmatic set of components. Although a change in core affect may seem necessary for all emotional experiences it is possible that even this aspect of an emotional prototype need not exist for all emotional experiences to count.

Russell analogizes emotional prototypes to the value of certain arrangements of cards in poker. While there are very many different possible combinations of hands in the game, only certain ones have value. A five card hand containing only prime numbers for example is interesting and rare but useless in the context of the game. These arrangements and their values are not a priori truths about decks of cards but rather pre-agreed arrangements that allow for a certain useful social function to exist (playing poker):

In like manner, different societies have lexicalized somewhat different patterns in their emotion vocabulary. Even when card hands are dealt randomly, when a series of numbers is random, or when a series of basketball hits and misses is random, observers perceive patterns. In emotional life, the mixtures of emotions’ components are not random, but even if they were, people would likely divide them into discrete categories on the basis of perceived patterns.42

At the level of actual emotional experience things do not actually hang together

---

42 Russell 2003 (166)
perfectly and the analogy breaks down. Emotional meta-experiences, the act of judging that your current emotional experience is a token of type x, imposes structure onto our messy emotional lives.

One advantage of a view like Russell's or de Sousa's is that it accords a proper respect to the various contributions that apparently competing views of emotion make. By making a distinction between emotional episodes and emotional prototypes we can see that some theories that claimed to be giving an analysis of emotional episodes are in reality accounting for an aspect of our emotional prototypes. Physiological perceptual theories like William James' for example are better understood as giving an account of the perception of affective quality and perhaps of emotional meta-experience. Cognitive theorists of emotions stressed meta-cognitive judgments and dimensional appraisals. Both theories are thus compatible with one another at the level of prototypes. Apparent counterexamples to each view thus target actual emotional episodes but this is to make a kind of category mistake. Russell’s theory is thus in a sense more inclusive than even the all-components necessary theories and can serve to explain both the state of the current debate along with our actual emotional experiences. Agreeing with him that emotions are complex and likely not natural kinds can therefore accommodate de Sousa’s original insight that the problem of emotional theory is in mapping the various aspects of those experiences. A prototype theory of the emotions is poised to address the intricacies involved in mapping different emotions without the pitfalls encountered when assuming that all emotional episodes
themselves share a set of necessary and sufficient features.

**Conclusion**

Conceptions of emotion often make the mistake of focusing on *one* of the elements in de Sousa’s analysis and calling that one element The Emotion. These theories often fail to capture important aspects of our actual lived emotional experiences and as such impose a greater burden on themselves to account for the difference between the theory of emotional experience and emotions as they are experienced. Every possible aspect of emotional experience has been singled out: perceptions, beliefs, desires, feelings, behavioral dispositions, physiological changes, judgments *about* emotional experience, and the cognitive effects of emotional experience (i.e. depressive realism). These conceptions, I argued, are unable to adequately account for our actual range of emotional experiences. ‘Hybrid’ accounts of emotion attempt to combine various aspects of emotional experience, usually in terms of some combination of beliefs, desires, and feelings. Of the available hybrid accounts, I argued that a view like Psychological Constructivism best “maps” onto important features of our actual lived emotional lives.

Russell successfully carves a middle position between those that claim that emotions are biologically based and those that claim that emotions are socially constructed. He is not alone in giving this kind of account but I believe his account, or an account like his, best explains the data from empirical psychologically while staying faithful to our everyday experiences with the emotions.
The following chapter will take up the issue of emotions within the context of new developments within Sentimentalist moral theory and assess their compatibility with the kinds of questions about emotion that have occupied us here.
Chapter 3

Getting a Feel for Sentimentalism

That emotions matter in ethics is uncontroversial. Being clearer about how they matter helps identify important points of disagreement in meta-ethics. I have been following convention by calling theories of value that give pride of place to the role of emotion 'Sentimentalist' theories of value. These Sentimentalist theories of value have undergone significant development in the last decade. It remains to be seen whether these emerging forms of Sentimentalism can explain important features of evaluative discourse. One of my aims in this chapter is to explore the degree to which these developing Sentimentalisms can explain moral discourse. Doing this requires getting clearer about the different ways that emotions have been incorporated into theories of value and it requires specifying what commitment makes a theory of value distinctly Sentimentalist in flavor. This commitment is to what I have called a kind of constitutive relationship between emotions and values. I expand on how we should understand this relationship. There are different ways in which a Sentimentalist can commit herself to this constitutive thesis. I will argue that contemporary Sentimentalists disagree on whether a first or second-order commitment to the constitutive thesis is best. second-order Sentimentalism, a Sentimentalism where normativity is determined by judgments about emotions and not by occurrent emotions, shows the most promise but neither first or second-order iterations can adequately explain the structure of our evaluative disagreements.
3.1 Emotional Roles

Kant is perhaps most famous for arguing that it is a grave mistake to use our emotions as guides for moral behavior. He argued that emotions are as likely to lead us to act immorally as they are to lead us to act morally. Our compassion may lead us to act rightly if it drives us to help an injured person on the side of the road. Kant cautions us that this very same compassion, however, could lead us to cover up for the wrongdoing of our children or loved ones. Actions done from an emotional basis, according to Kant, do not demonstrate a goodwill for just this reason. Instead Kant sought a more secure, in his eyes a more stable, source for our moral judgments. Kant reasoned that only a will that acts in accordance with duty for the sake of duty itself had value and he went on to give an analysis of duty that was based on the distinctly human capacity to rationally examine and set our own maxims. This is the capacity that helps us pick the grounds for action that will always lead us to do our duty. Some maxims for actions contradict themselves when universalized either inherently—one might say conceptually—or psychologically as when we think we can survive without the aid of others. Emotions, on this view, corrupt our will. Emotions drive us to act for the sake of something other than duty itself.

Sentimentalists place themselves in opposition to the rationalist conception of the role of emotions. A rough attempt at a definition of Sentimentalism is that Sentimentalists believe that emotions are necessary for moral judgments or actions. This creates some distance with, for example, some readings of Kantian moral theory;
in particular it differentiates itself from those rationalist views that stress that acting in accordance with reason against our inclinations—Kant's term for our emotions—has moral worth whereas acting in accordance with reason because of, say, sympathy, lacks moral worth. But simply saying that emotions are necessary for moral judgment does not clarify what it is about the emotions that a Sentimentalist takes to be important.

Sentimentalism implies a commitment to the claim that emotions are playing a necessary role in our moral or evaluative worlds. This claim may actually imply that emotions are playing at least one of the different roles described below:

**Moral Motivation**: emotions are necessary for moral motivation. Without our emotions we would not be motivated to behave in accordance with the demands of morality or with our moral judgments or beliefs.

**Moral Epistemology**: emotions are necessary in order to understand moral concepts and evaluative properties. Without the emotions we would not be able to accurately track moral truths and hence fail to recognize moral demands.

**Moral Constitution**: the content of emotional responses is necessary in order to explain our moral or evaluative concepts. It is impossible to explain what our moral or evaluative concepts mean without mentioning our emotional responses.

It is important to be clear what each thesis entails about the relationship between emotions and morality and whether a commitment to any of these theses implies Sentimentalism. The motivational thesis is perhaps the most obvious role for emotions. Our moral judgments, it is normally conjectured, seem to motivate us

---

1 Moral worth may be viewed as a kind of supererogatory merit and not as a requirement for right action.
(barring weakness of will) to do something. Suppose you see two people in a heated argument. On closer inspection you realize the argument is actually a robbery. You judge—or hopefully already have the standing belief—that robbery is wrong. This judgment, it is often assumed, cannot be made idly. Judging that something is wrong stirs you to action by invoking moral outrage toward the perpetrator. In this case it might stir you to call the police or in some other way attempt to stop the mugging. The view that moral judgments necessarily motivate is called motivational internalism. The motivational thesis thus might best be seen as an explanation for the internalist intuition. If affective states (including emotions or desires) were coupled with moral judgments then it becomes easier to explain why it seems like all moral judgments seem to have the property of motivating us to act. Although Sentimentalists often make a commitment to the motivational thesis, I argue that it is not by itself an essential feature of Sentimentalism. Sentimentalism, I argue below, is a thesis about the grounding or meaning of our moral claims. The motivational thesis, on the other hand, is a claim about human psychology. Since what we are after is a distinctly Sentimentalist claim (if there is one to be had) then this will not do. It is entirely possible to believe that moral claims are principles of (pure) reason and that emotions are co-opted only to serve motivational purposes that get us to do the right thing. This would connect a form of rationalist meta-ethics with the motivational thesis and serves to show that motivational internalism is separable from Sentimentalism.

What of the epistemological thesis? On the epistemological thesis emotions are
like speedometers. Speedometers are reliable indicators of velocity insofar as we can track real-time changes in velocity by gauging the movement of an indicator arrow on our dashboard speedometer. Our speedometer's arrow moving from 40 to 50 is correlated with a change in velocity. The relationship between velocity and arrow location is uni-directional however. When properly calibrated, only real changes in velocity cause our speedometer's arrow to move. Manually moving the arrow on a speedometer does not affect our velocity. Velocity, we might say, is independent from the speedometer. Speedometers serve as a useful heuristic for us to keep track of something we independently care about: velocity. Emotions, on the epistemological thesis, function in just this way: they reliably track moral values. Suppose again that you are a witness to a robbery. A deep sense of outrage stirs noticeably within you but beyond the outrage you also notice feelings of guilt at the thought of not doing anything to change the situation. According to the epistemological thesis your feelings of outrage are cluing you in on the fact that an injustice is occurring. The feelings of guilt that you feel are serving to make you aware of a duty to act to stop injustice. In this case injustice and duty should be understood as independent of the emotions themselves. An injustice would be occurring whether or not you felt outrage and you would still have a duty to do something to stop that injustice even if you did not feel guilty at the thought of inaction. As in the speedometer case, the emotions are functioning as heuristic tools. They track moral truths (like those related to Justice) but they do not cause something to become moral—just as manually moving the
speedometer's arrow does not cause a change velocity. The epistemological role is also not essential to the Sentimentalist theory of value. It is possible for a rationalist (committed to a particular view of human psychology) to co-opt the epistemological thesis.² On such a view our moral duties might be grounded on claims about what kinds of actions we might not be able to rationally reject. This would give morality a clearly emotion-independent basis. Our moral emotions, on this kind of rationalist view, would serve as indicators of the kinds of things or actions that are not rationally rejectable. If Sentimentalists are to be committed to a distinct set of claims about the nature of morality or values then they need to claim that emotions are doing something more than just tracking our values if they want to avoid the kind of rationalist co-opting strategy I have been making use of thus far. The emotions must somehow play a role in our moral judgments that the rationalist cannot simply add on to their own theory. They must somehow argue that emotions constitute moral concepts such that moral concepts are not moral unless our emotions make them so. This needs clarification.

The constitutive thesis, it seems, is the essence of Sentimentalism. The thesis is a claim about the grounding of our moral concepts; it commits one to the view that the content of a suitable set of emotional responses explains—that is, can account for—the content of at least some of our moral concepts. Unlike the epistemological thesis where values are independent from our emotions, the constitutive thesis claims that at

² The commitment in mind is a rationalist that believes that moral facts are principles of reason but whom also believes that humans can best access those truths via our emotional responses because emotions (perhaps for evolutionary reasons) reliably track those values.
least some values are valuable because of their relationship with the sentiments. On
the epistemological thesis a uni-directional relationship exists between our values and
our emotions. On that view evaluative properties normally tend to cause us to respond
with emotions. The constitutive thesis requires this relationship to be bi-directional
and for the role of emotions to be ineliminable from an analysis of moral concepts.
For example some Sentimentalist theories of moral responsibility argue that our
understanding of moral responsibility should be understood in terms of who we think
it is appropriate to feel resentment, indignation, or guilt toward. Any analysis that
explains moral responsibility that does not address the function of our emotions
would, on this view, be incomplete. The claim then is that some emotions are such that
they not only track value (the epistemological thesis) but that they make something
valuable. Value is essentially tied to the content of our emotional responses.

3.2 Two kinds of Sentimentalism

If Sentimentalism is a distinct view of the relation between morality and
emotion then that distinction lies in the commitment to the constitutive thesis. I argue
that Sentimentalists have construed a commitment to this thesis in different ways:
first-order and second-order. I also argue that both views ultimately have problems

---

3 This might seem to miss the mark. Settling who is responsible may not tell us why someone is
responsible. Peter Strawson and Jay Wallace however, have argued that an analysis of the conditions
under which our emotions are appropriate will give us an account of what responsibility itself
consists in. See Peter Strawson “Freedom and Resentment” (Proceedings of the British Academy
Vol 47, 1962) 1-25 and Jay Wallace Responsibility and the Moral Sentiments (Cambridge, Mass:
Harvard University Press 1994). Wallace and Strawson's theories of moral responsibility will be
the focus of the sixth chapter.

4 I go into greater detail on the nature of the constitutive thesis in chapter five in the section on
response-dependent theories of value.
explaining evaluative discourse. first-order Sentimentalism grounds the normativity of emotions by appeal to the actual emotions that an agent would have under some set of, usually idealized, conditions. David Hume, for example, famously wrote that

> When we form our judgments of persons, merely from the tendency of their characters to our own benefit, or to that of our friends, we find so many contradictions to our sentiments in society and conversation, and such an uncertainty from the incessant changes of our situation, that we seek some other standard of merit and demerit, which may not admit of so great variation. Being thus loosen’d from our first station, we cannot afterwards fix ourselves so commodiously by any means as by a sympathy with those, who have a commerce with the person we consider…we blame equally a bad action, which we read of in history, with one perform’d in our neighborhood the other day.\(^5\)

Here Hume notes that our actual emotional responses are likely to be inconsistent and biased but that our moral principles are not inconsistent or biased.\(^6\) This problem, I will call it Hume’s Problem, needs to be addressed by any aspiring form of Sentimentalism. Because of Hume’s Problem, our actual emotional responses cannot serve as the ultimate normative ground for virtue if we want to keep the belief that virtue involves a consistent set of demands or actions. In response to this problem, Hume goes on to develop an impartial or idealized moral standpoint. From this impartial standpoint our emotions would be consistent and unbiased and hence the definition of virtue that arises from this standpoint would retain those features. This is why we know, according to Hume, “from reflection” that wrongdoings in the past are just as wrong as wrongdoings being committed today even though they both, as a

---


\(^6\) Not all Sentimentalists follow Hume here. Some Sentimentalists are committed to a form of moral particularism that denies the existence of codifiable moral principles. I say more about Sentimentalist particularism in chapter five.
matter of fact, are unlikely to stir the same emotions within us to the same degree of liveliness. An impartial spectator would have the same emotional responses to both situations and hence the moral standing of identical wrong actions separated only by distance or time would, from her point of view, be the same. The first-order approach fixes the content of moral principles on the emotional responses of impartial or otherwise idealized spectators. This view is first-order because the actual emotional responses of anyone taking the impartial standpoint are sufficient to ground the moral nature of the state of affairs being judged. There are other ways of construing first-order Sentimentalism. Jesse Prinz, for example, has recently defended a version of first-order Sentimentalism that is explicitly relativist in its scope. On a first-order view of this kind our actual non-idealized emotions are enough to ground the moral standing of a state of affairs. Because of Hume's Problem a view like Prinz's would limit the scope of moral claims only to those who share the same kinds of responses as our own. For many this results in unpalatable moral relativism. Here then are two examples of a first-order Sentimentalism that take different approaches regarding the scope of our moral claims. For Hume the impartial standpoint will lead us to make relatively universal moral judgments once we are stripped of potential emotional biases. For Prinz our actual, biased, emotional responses serve to delineate the moral community such that only those that share these responses are addressable by our moral claims. Problems arise for the first-order view when one tries to justify the

---

7 Ibid.
8 More recently, Roderick Firth developed the idealized observer theory in his “Ethical Absolutism and the Ideal Observer” Philosophy and Phenomenological Research 12, 1952: 317-345.
conditions of idealization without begging the question. Advocates of partiality, for example, would resist making full impartiality a feature of the ideal observer on the grounds that some duties are rightly partial (those owed to family for example). Furthermore, we might come to question why we should feel any obligation to do what an idealized or impartial spectator would do. The conditions of idealization may so dramatically change the nature of idealized agents that their judgments can fail to find purchase with us. We may also be led away from a Prinzian first-order Sentimentalism due to the sting of Hume's Problem. If what we want is to be able to make moral claims that apply not only to those that agree with our moral claims or who share our more general moral sensibilities but to those with whom we disagree then a view like Prinz's will not do. Since this kind of first-order theory cannot explain how moral address of this kind is possible we may be tempted to move away from first-order Sentimentalism more generally.

The second-order approach avoids many of these problems. second-order Sentimentalism is currently the favored way of explaining the normativity of emotion within the Sentimentalist literature. To get around Hume’s problem—a worry about the danger of extreme subjectivism—the second-order Sentimentalist seeks to ground the normativity of moral concepts not by appealing to the emotional responses of an

---

impartial spectator or of our emotional responses as we actually have them but instead by appealing only to those emotional responses that are appropriate or warranted or rational. Broadly construed, appropriateness can be understood as a kind of match between an emotional response and the world. The approach is second-order because our actual emotional responses are not constitutive of moral concepts but instead only appropriate emotional responses perform this function. This leaves much of the explanatory work to an account of what makes an emotion appropriate. There must be some appeal to emotional responses in this account. If, for example, appropriate emotional responses attached themselves only to rationally universalizable maxims then the appropriate emotions would be serving an epistemological role and Hume's Problem would be bypassed at the cost of Sentimentalism. When we feel an emotion toward an object that falls outside of what is sometimes called its locus of concern—whatever it is that emotions are 'about'—we may rightly claim that something has gone wrong, that the emotional response is inappropriate.

There is another wrinkle that crops up against second-order Sentimentalism. Jesse Prinz has recently attempted to argue that second-order Sentimentalism 'uses emotions in the wrong way' and hence fails to adequately capture the Sentimentalist intuition. Although his main targets were John McDowell and David Wiggins, the critique applies more generally. McDowell and Wiggins endorse what Prinz calls a meta-cognitive theory of moral concepts. This means that moral concepts are comprised not out of just any emotional response but out of appropriate emotional
responses. For this reason Prinz's meta-cognitive theory of moral concepts is identical to what I have been calling second-order Sentimentalism. Prinz calls these views meta-cognitive because they require an evaluative attitude about our emotions in order to understand the content of our moral concepts. It seems appropriate here to note that one of the advantages typically supposed by meta-cognitive theorists of emotion is that they are able to account for substantive moral disagreement within a subjectivist meta-ethical framework. This advantage is significant and, as I will have reason to note in a moment, Prinz's own view, while admirable and ingenious in many respects, has a problem accounting for this particular kind of moral disagreement.

Prinz’s theory is not meta-cognitive. His approach is first-order. Moral concepts are comprised of sentiments (not appropriate sentiments or warranted sentiments but sentiments simpliciter). Though there is some space here between the sentiments as emotional dispositions and the emotions themselves. Prinz places the normativity of his theory on the sentiments. Thus, merely having an emotional response to some x may not make it a moral response even if the emotion we feel towards x is one of the moral emotions, I would think, unless we experience that emotional response as a result of having already had a standing disposition (a sentiment) to feel that emotion about things like x. So long as this is true then the simple possession of the sentiment is enough to say that the agent has the corresponding moral attitude toward x. This discussion, I think, warrants some

---

11 This chapter has in part sketched out, and critiqued one method of doing just this: the Rational Sentimentalist's method due to its reliance on basic emotions.
questioning.

Prinz's chief complaint against the meta-cognitivists is that they fail to use emotions 'in the right way' when they ground moral value on appropriate emotional responses. In his terms, moral judgments mention emotions but do not use them. This claim is unclear. A sophisticated meta-cognitive Sentimentalist would argue that our conceptions of appropriateness are drawn from a phenomenological analysis of the emotions themselves in order to see what it is that they are about, what they are concerned with. A phenomenological analysis of sadness might lead us to conclude, in agreement with Prinz, that sadness seems to be a response to some kinds of losses. Having gone this far the meta-cognitive takes a further step by giving an account of what the central concern for sadness might be. A rough gloss on behalf of the meta-cognitivist might be that sadness is concerned with the loss of something significant to us. Having given this kind of gloss the meta-cognitive can then argue that only instances of sadness that are responses to losses of this kind are appropriately lamentable. That is, only those losses are appropriate targets for sadness. The concept of appropriateness is here tied very strongly to the function of the emotion in question. Appropriate emotions are those that get something very specifically right about the relation between the subject of the emotional experience and the object of the emotional experience. Here then the emotions themselves, when appropriate, delineate different realms of value as opposed to Prinz's view where larger emotions dispositions (the sentiments) ground values.
On the face of it this does not seem very different from Prinz's own analysis of how emotions can come to be rationally assessable. In his terminology the meta-cognitive Sentimentalist is merely making the claim that only correct emotions—emotions that are correctly attributing a property to something—are constitutive of moral properties. The feelings are just as essential on this view as on Prinz's view, I would argue, since both have emotions at one degree of removal from moral properties. In Prinz's case the emotions themselves are not the bearers of moral properties, the moral sentiments are. In the meta-cognitive case it is not just any emotional response that counts as a bearer of moral properties but only those that get something right about the relation between the emotion, the agent, and the world.

The problem with appropriateness, according to Prinz, is that something can be appropriate or inappropriate in a number of different ways and some are more problematic than others. For example, what kind of appropriateness is, well, appropriate for moral properties? Prinz's answer is that moral appropriateness seems like the right response to the question but that all that response does is introduce a kind of useless circularity to the definition of rightness (wrongness). However, I believe this is a weak reason to reject meta-cognitivism especially considering its ability to account for cases of disagreement about value.

Prinz takes this problem—noteing Justin D’Arms and Dan Jacobson as the first to make the exact nature of it clear—to be a major cost of the meta-cognitive approach. He thinks, again, that adopting the first-order view gets around this
problem. Meta-cognitivists have a response to this worry however and Prinz would do well to at least give some reason for thinking that their response is inadequate. According to D'Arms and Jacobson, emotional appropriateness is not the same as pragmatic appropriateness or appropriateness based on rules of etiquette or even moral appropriateness. Instead emotions are appropriate iff that emotion is responding accurately to its locus of concern. This is a kind of appropriateness that is uniquely emotional. It is grounded on our experience with what our emotional responses seem to be about, and hence can get around the problem of having too many standards of appropriateness. For example D'arms and Jacobson argue that a joke could be genuinely funny (and hence that finding the joke amusing is appropriate) even if it is in bad taste (against rules of etiquette) or pragmatically bad for us to tell (because people would think we were insensitive or bigoted). 12

On this score then I find that Prinz's first-order Sentimentalism and the meta-cognitivist's second-order Sentimentalism to be on all fours. Both can account for moral properties and can account for mistaken instances of emotional responses. In Prinz's case the moral properties are tied to emotional dispositions (emotions in the approbation and disapprobation spectrum) while the meta-cognitive finds moral value in having appropriate moral emotions. Both have actual emotional responses at one degree of removal from moral properties.

In what follows, the analysis of second-order Sentimentalism will focus

12 See Jesse Prinz's The Emotional Construction of Morals. (Oxford: Oxford University Press, 2007) especially Ch. 6 “Dining with Cannibals.”
specifically on ‘Rational Sentimentalism,’ a second-order Sentimentalism under development that assigns emotions a uniquely rational function.

### 3.3 Rational Sentimentalism

Rational Sentimentalism is a Sentimentalism in which emotions are subject to a specific kind of rational assessment. As a second-order Sentimentalism the nature of evaluative properties is dependent only on emotions that survive this assessment in terms of appropriate emotional responses. Only appropriate emotional responses, on this view, constitute the content of our evaluative concepts. Evaluative concepts like the funny cannot be understood without appealing to the kinds of things that make human beings laugh. Evaluative properties must make reference to subjective human responses in order to count as an accurate explanation of the meaning of the property.\(^{13}\) Furthermore what it means for a joke (an object) to be funny (an evaluative property) is for it to be appropriate for a properly situated agent to laugh at that joke. Being properly situated does not necessarily imply that Rational Sentimentalists appeal to idealized or impartial spectators.\(^{14}\) These responses set the initial boundaries of the concept under analysis. This is necessary because, the Rational Sentimentalists argue, the kinds of things that evoke our emotions do not share a set of necessary and sufficient features that allow them to be classified in an observer-independent objective fashion:

---

\(^{13}\) It isn’t clear whether the Rational Sentimentalists wish to commitment themselves to the empirical claim that all (relevant) emotions have visible or identifiable behavioral responses like laughter.

\(^{14}\) To be properly situated to laugh at a joke, for example, one must have the relevant background information to ‘get’ the joke, be able to physically laugh, etc. These minimal conditions are all that is implied by ‘properly situated.’
We could say that what makes something funny, for instance, is its incongruity; and what makes something fearsome is its dangerousness —where incongruity and danger are understood as empirical, observer independent properties. However, our patterns of response and criticism are too unruly for any such treatment to be tenable and substantive. If our responses are uncodifiable, and if values depend on those responses, then values too cannot be codified in rules.\textsuperscript{15}

Attempts to categorize the content of our evaluative concepts in an observer-independent fashion, like reducing 'the Funny' to 'the incongruous,' are bound to fail because the concept under analysis is mind-dependent. The Rational Sentimentalists also argue that a simple commitment to the epistemological thesis runs counter to our experience with emotion. This is because our emotional responses do not simply track evaluative concepts but constitute them and the evidence for this is, in part, because it is not possible to reduce the content of our evaluative concepts to a set of observer independent properties.\textsuperscript{16} The only way we can fully understand the Sentimentalist evaluative concepts is via reference to appropriate emotional responses.

The sentiments may be unruly but they do not respond at random to any object or state of affairs. Because the Rational Sentimentalist “aspires to vindicate the phenomenology of valuing as a matter of sensitivity to features of the world, while acknowledging that values are founded on human sentimental responses” emotions are

\textsuperscript{15} See Justin D'Arms (2005) p. 12-13
\textsuperscript{16} The Rational Sentimentalist's arguments for the constitutive thesis will be examined in more detail in chapter five of the dissertation. Although we are able to give definitions for the locus of concern for an emotion, Rational Sentimentalists argue that it is unlikely, due to the unruliness of emotional responses, that any particular articulation of the locus of concern for an emotion will be definitive. Thus the locus of concern for an evaluative property is essentially contestable: it is impossible to settle once and for all because emotions, on the Rational Sentimentalist view, lack constitutive thoughts and hence the project of determining the locus of concern is partially a phenomenological one.
described as having evolved to pick out sets of features out in the world.\footnote{See Justin D'Arms (2005) p. 12-13} When we are in the grip of an emotional experience, how does the emotion present the object to us? When we are envious of our neighbors’ new car what is it about how we experience this emotion that presents the car as especially enviable? D’Arms, for example, has argued that a distinctive feature of envy is that it serves to indicate the loss of access to or possession of positional goods.\footnote{See Justin D’Arms' Stanford Encyclopedia entry on Envy available at: (http://plato.stanford.edu/entries/envy/). For the Rational Sentimentalist stating the locus of concern for envy doesn't amount to a reductive/non-emotional theory of the enviable. Not all losses of positional goods are genuinely enviable. The contours of the enviable are not reducible to the observer-independent property of positional goods.} The basis for this claim is, at least in part, an analysis of the way that enviable things seem to us when we are, in fact, feeling envious. Appropriate emotions respond only to their locus of concern and we figure out the locus of concern by building up experience with the emotion and its evolutionary function.

Having a sense of the locus of concern for an emotion lets us talk about specific kinds of reasons that might speak in favor of feeling (or not feeling) a particular emotion. Not all reasons function equally well. Reasons that speak to the fittingness of an emotional response are independent from other kinds of evaluations of the same emotion: that it is imprudent or immoral to feel. It may be immoral to be envious of your neighbor’s success but the fact of the immorality (if it is a fact) does not make that success any less enviable. The evaluative property consists of the particular way in which emotions present objects as being when we are in the grip of
the emotion, in this case what it means to say that something is enviable.\textsuperscript{19} On this view, it would be appropriate to envy your neighbor, and your neighbor's success would be enviable, whether or not you \textit{actually} ever feel envious. Evaluative properties are independent from any particular response that an agent might have. It is because reasons of fit speak to an emotional response’s locus of concern and because the locus of concern is identifiable independently from any particular agent (but not independent of any agents whatsoever) that this degree of independence is possible.

A judgment that an emotion is fitting requires the ability to both \textit{feel} and \textit{regulate} the response.\textsuperscript{20} Rational Sentimentalists argue that, as a matter of \textit{fact}, the judgment that an emotional response is unfitting is likely to extinguish or at least reduce the emotional response in question (and vice-versa for fitting judgments) and it will do this more successfully than other ‘external’ pragmatic considerations. If an agent possesses basic emotional capacities then judgments of fittingness will have \textit{normative} force for those agents: they will play a regulative role by giving them \textit{reasons to feel}. This ‘regulative role’ argument helps us to see how our emotional responses function to give \textit{reasons}, albeit reasons of a particular sort, to agents. It is because of the combination of the Rational Sentimentalist's theory of emotions and the regulative role argument that the Rational Sentimentalists set themselves apart from other Sentimentalist theories. Were we not sensitive to the appropriateness conditions for the funny, we would not be able to fully understand the reasons that make

\textsuperscript{19} See Justin D'Arms and Daniel Jacobson's “Sentiment and Value” \textit{Ethics} 110, 2000 p. 722-48

\textsuperscript{20} See Justin D'Arms's “Two Arguments for Sentimentalism” \textit{Philosophical Issues}, 15, Normativity, 2005
something funny and hence we would fail to fully understand the evaluative concept.

The Rational Sentimentalists get around Hume’s problem—that our emotions can be inconsistent and biased even though our moral principles are not—by appeal to their specific notion of emotional appropriateness. The concepts delineated by our appropriate emotional responses achieve consistency in a way that our first-order emotional responses do not. Since only appropriate emotional responses are constitutive of evaluative properties then Hume's Problem is addressed. Our values can be consistent even if our emotional lives are not.\(^{21}\)

They also address Jesse Prinz’s Problem of Plenty, outlined in the previous chapter, via the same theory of emotion. If full-blown emotional concepts are the product of evolution then evolutionary processes also determine the structure of an emotion’s component parts and those parts will be structured, it is assumed, in whatever way was most fitness enhancing.\(^{22}\)

To summarize: Reasons of emotional fit are particular. That an emotional response is fitting provides any human being capable of feeling and regulating the emotion with a defeasible reason to feel it. Reasons of fit serve a regulative function and bring emotional responses into the space of giving reasons by making them assessable to a particular sort of rational evaluation.

Second-order theories like Rational Sentimentalism represent the current state of the art in terms of Sentimentalist theories of value. Although advances are made in

\(^{21}\) Note that the first-order Sentimentalist has a choice when it comes to Hume's Problem: deny it is a problem by stating that morals are relative or adopt a 'basic' emotions strategy such that we should expect a degree of agreement about our moral responses given their nature.

\(^{22}\) Basic emotions are discussed more fully in the proceeding chapter.
terms of addressing Hume's Problem and in providing an alternative to the relativism of the first-order views, significant problems remain.

3.4 Disagreement

George and Frank meet up for coffee after watching *Night of the Living Dead.*

George loved the movie and Frank cannot understand why. They engage in the following conversation:

George: It was fantastic! The film had me on the edge of my seat.

Frank: I don't get it Frank. The zombies weren't scary, they were too slow.

George: They're supposed to be slow. Zombies aren't scary because they hunt you down, they're scary because they're relentless! They never sleep, they never get tired, they're always there and always hungry.

Frank: That might be true but the threat just isn't imminent. Zombies are still too slow and there weren't enough of them in the film to be imminent threats. The posse at the end of the film, they were a serious threat and I think showing them undercut my fear of the zombies.

George: That's the point! The zombie threat forces everyone into a primal survival state. They end up being the source of their own demise by becoming like the monsters they were running away from! It's brilliant.

Frank: If that's what the film was trying to say then I think it's just not an original point to make. Too many other films have done that to make it interesting enough to justify the film. No wonder I had a hard time finishing it!

Conversations like this are the bread and butter of our argumentative lives. The disagreement is evaluative because the subject of the disagreement is whether or not *Night of the Living Dead* is a good horror film. As a contrast, a non-evaluative disagreement about the same film might be whether the director is male or female.

who the wrote the score for the film, or whether it was an actor's debut film; in other words, evaluative disagreements are not about facts but about values. George and Frank disagree about several things. In part their disagreement is about the nature of fear and whether zombies are fearsome. In addition they disagree on whether the film itself is a good horror film. In both cases Frank and George take themselves to be offering reasons they expect one another to understand and potentially accept. They do not see their conversation as one where they are merely offering their opinions. They are trying to change each others minds. Any theory that purports to explain our evaluative concepts must be able to explain (or adequately explain away) evaluative disagreements like the one above (and many countless others).

Parties engaged in an evaluative disagreement see themselves as trying to settle a question about the right way to think or feel about something. For some evaluative judgments we are content to let our taste determine an answer but cases like these are not disagreements at all. We are normally fine leaving it up to personal taste whether rocky road or chocolate chunk is the better ice cream flavor. I could agree that you are a chocolate chunk lover and you could agree that rocky road is best for me. Evaluative disagreements have a different character. Evaluative disagreements consist largely of appeals to a set of criteria that can be used by all parties to evaluate an object. Frank and George disagree on whether the features of Night of the Living Dead merit classifying it as a good horror film. They both seem to have some shared sense of what a good horror film should be like but they disagree with one another on whether Night
of the Living Dead should be considered one. It becomes a challenge for Sentimentalists to explain how disagreements like this can happen if our emotional responses are what ground our value judgments.

In order to see why suppose we’re naïve first-order Sentimentalists. We believe that evaluative concepts are grounded on our actual emotional responses. With an extreme Sentimentalism like this the problem of evaluative disagreement becomes clear. If two individuals disagree about whether a film is a good horror then what kinds of reasons can they bring to bear in order to change one another’s minds? It begins to look like the ice cream case. For example, George may claim that genuinely fearsome things necessarily need to invoke the unknown while Frank may claim genuinely fearsome things need only be threatening. If emotions are idiosyncratic and if evaluative concepts are grounded on those emotions, then there is not much George and Frank can say to one another that would constitute a genuine disagreement about fearsomeness. George can come to agree that Frank finds things fearsome when they are threats and Frank can agree that George finds things fearsome when they invoke the unknown. They fail to actually disagree. In fact it is worse than this. They actually agree with one another. If evaluative disagreement requires an observer-independent set of criteria in order to disagree about values then Sentimentalism appears to deny that values can be observer-independent. Moral concepts, furthermore, are often believed to be objective and their being objective easily explains how it is possible that subjects could debate about them. Because Sentimentalists ground moral concepts
on emotional responses they need to give an account of those emotional responses that can ensure that agents engage in genuine evaluative disagreement while at the same time explaining why it seems like moral concepts are independent from our particular emotional responses.

Rational Sentimentalists believe they have a ready reply to this kind of problem which I rehearse here and then take up more fully in the next chapter.

Emotions, recall, are evolutionarily derived reflexive modules. They can take features of the environment relevant to their loci of concern as input and produce emotionally salient thoughts, judgments, and patterns of behavior as output. These patterns of thoughts, judgments, and behavior are further thought to be identifiable with emotion specific patterns of physiological change. Because emotions are conceived of as basic—that is, phylogenetically primitive—they are often thought to be present in most mammalian species. The upshot of this for the Rational Sentimentalist is that it becomes objectively verifiable whether a genuine disagreement is happening between two agents. If two agents can be shown to be in the same emotional state (that is, to be grounding their evaluative claims on the same emotional response) then they are disagreeing about whether or not something should appropriately fall under its evaluative concept.\(^{24}\) For example, Noel Carroll has characterized the emotion induced by horror film and literature as 'art horror.'\(^{25}\) If art horror is indeed an emotion—and if not then fall back on garden variety horror or fear or disgust—then it should be

\(^{24}\) I've alluded to the fact that D'Arms and Jacobson believe that emotions can be individuated by purely physical markers and hence could be objectively measured. The following chapter will discuss these issues directly.

\(^{25}\) See Noel Carroll's *The Philosophy of Horror: Or, Paradoaxes of the Heart* (Boston: Routledge 1990)
verifiable whether Frank and George were arguing over whether Night of the Living Dead merits our responding to it with art horror. We could track their heart rates, pulses, eye blink rates, or galvanic skin responses and see whether their arguments are grounded on the same pattern of physiological change. If so then a genuine disagreement about the emotionally grounded evaluative concept 'thrilling' is happening.

This account is overly optimistic. Emotions, it seems, are not verifiable in the way imagined above. I here only briefly touch on the data regarding basic emotions. The following chapter will devote itself to arguing more fully against tying Sentimentalism to basic emotions if only because the data in support of basic emotions is in doubt. Although the belief in emotionally specific patterns of physiological or neurological change is an old one, there has yet to be widely duplicated experiments in which these patterns reveal themselves. Activation of particular regions of the amygdala for example were once thought to be specific to fear though the amygdala is now thought to mediate different kinds of positive and negative affective stimuli. Only gross affective changes (entering positive or negative emotional states for

---


example) have been found to be predictable from physiological or neurological data. Furthermore evidence is building that suggests that many early studies thought to support basic emotions are weaker than supposed. These are known as “judgment studies.” In a typical judgment study subjects are shown a facial expression and asked to pick out which emotion (from a list) it expresses. These studies tended to have high rates of agreement across cultures. These studies have been critiqued on both methodological and conceptual grounds. I say more about these judgment studies and the critiques of them in the following chapter. For now I will say that it was often thought that the kind of cross-cultural convergence that judgment studies appear to offer was best explained by the existence of basic emotions but this inference is no longer sustainable. If true then this adds serious doubt to whether the version of Sentimentalism developed by D'arms and Jacobson can adequately address the kind of everyday evaluative disagreements under consideration. Although this need not spell doom for second-order Sentimentalism more generally it does pose a serious threat to Sentimentalist theories that understand emotions as basic. If basic emotions are the only route available to Sentimentalists to explain how evaluative disagreement is


possible then the view has a serious empirical hurdle to overcome.

Conclusion

Sentimentalism is best understood as offering a theory of the grounding of moral properties. Emotional responses constitute those properties. Constitution requires that an emotional response and moral property be ineliminably biconditionally related. It must be an ineliminable feature of a proper explanation of the meaning of an evaluative property that it is invoked by emotional responses. The constitution relation can be construed as first-order or second-order. First-order Sentimentalism has a hard time addressing Hume's Problem. Second-order Sentimentalism generally, and Rational Sentimentalism in particular, can address Hume's Problem by tying moral properties to a special class of emotional responses: appropriate ones. While this is an advance, both Prinzian first-order and D'arms and Jacobson style second-order Sentimentalism have trouble explaining how it is that evaluative disagreement is possible. Sentimentalists need to be able to address this problem before further progress can continue.

Rational Sentimentalism appeals to two unique claims about emotion. The regulative argument is separable from the commitment to basic emotions. In the following chapter I will take on the question of basic emotions directly and will argue that the evidence in favor of basic emotions is shaky and can be accounted for with a non-basic theory of emotion. Rational Sentimentalists argue that basic emotions are necessary in order to explain evaluative disagreements. I will argue that this is false,
up to a point. Basic emotions would secure a fairly robust scope for our moral
judgments, the alternative I propose would have a more limited scope but would,
within that scope, still be able to explain moral disagreements. Given the status of
basic emotions I suggest it is better to do without them and that a Sensible
Sentimentalism should be tailored to satisfy this requirement.
Chapter 4

Rational Sentimentalism and Basic Emotions

In characterizing Rational Sentimentalism I said that evaluative concepts were
dependent upon emotional responses because the content of the evaluative concept are
at least partially constituted by the way in which an object or state of affairs is
presented to us when we are in the grip of an emotional episode. In earlier chapters I
called this the constitutive thesis. I also said that Rational Sentimentalists make a
special kind of appeal to reasons that speak to what they called the fittingness of
emotional responses. A fitting emotional response is, on this view, based only on
whether the emotional response correctly attributes to its object the relevant properties
with which that particular emotion is concerned. If envy is concerned with the loss of
positional goods then envying the loss of a genuine positional good would be a fitting
response in that circumstance. On the Rational Sentimentalist picture, the matter of
correctly attributing a property is dependent on the evolutionary basis for emotion.
This gives us a clear roadmap of the kind of moral epistemology that the Rational
Sentimentalist can offer. What did this basic emotion evolve to detect? Am I detecting
it now? If the response is accurate then it is fitting and if fitting then one has a genuine
reason to feel the emotion. This reason can then enter the space of giving reasons and
play a role in rational deliberation. Fitting emotional responses thereby render an
object properly fearsome (funny, shameful, etc). If the response inaccurately attributes
features to the object, then it is not fitting and we do not have a reason (of this kind) to
But this explanation of Rational Sentimentalism leaves an important element out of the equation. How do the Rational Sentimentalists understand the emotions that are being giving us reasons? In the previous chapter I alluded to the fact that Rational Sentimentalists make appeals to a conception of basic emotions in their explanation of evaluative disagreements. There I said that they argue that unless we understand at least some emotions as basic we are lead to a potential problem: that evaluative properties and the reasons that speak to their fittingness may have a very limited scope. In fact, depending on how one characterizes the nature of emotional responses, the scope could be limited to the particular agent experiencing the emotion. If our emotions are idiosyncratic then the reasons these emotions would provide would also be idiosyncratic. This form of Sentimentalism would result in a form of agent relativism. This would be the case if emotions were, for all I have said, idiosyncratic to particular agents. I have attributed a view of this kind to Jesse Prinz in previous chapters. Here I want to more carefully examine how one's theory of emotion and one's position on whether any emotions are basic can impact the status of the moral theory that is built from those emotions.

A further problem: depending on the theory of emotion under discussion it may be difficult to explain how rational disagreement is possible. It would be, to say the least, an unattractive conclusion for any theory of moral values and moral discourse if

---

1 Though we may, as already noted earlier, have many other kinds of reasons for wanting to feel (or avoid feeling) an emotion that may range from the pragmatic to the moral.
it were unable to account for as basic a feature of moral discourse as evaluative
disagreement. To their credit the Rational Sentimentalists explicitly conceive of
emotional responses in such a way as to avoid this problem of rational disagreement:

\[
\text{[E]motions are syndromes of cognitive, affective, motivational, and} \\
\text{behavioral changes, which arise in patterns displaying some degree of} \\
\text{consistency across times and cultures—perhaps because of our shared} \\
evolutionary history.}^2
\]

This characterization helps us to see how Rational Sentimentalists conceive of
emotion. However, it is still unclear what these cognitive, affective, motivational, and
behavioral components of emotions are. It is also unclear how emotions are
differentiated from one another and, most crucially, what exactly is holding constant
across times and cultures such that we can speak of the same emotions existing in all
of these cases. A bit more clarification is offered elsewhere though here the Rational
Sentimentalists call the emotions 'sentiments':

The sentiments are such paradigmatic types of emotion as amusement, anger, contempt, disgust, envy, fear, guilt, jealousy, joy, pride, regret, shame, and sorrow—which we believe will prove to be \textit{robust psychological kinds}...Many of these syndromes come together as suites of responses because they are, or are products of, relatively \textit{discrete psychological mechanisms} that \textit{evolved} for their adaptive value in dealing with “recurrent adaptive situations” (Tooby and Cosmides 1990) or “universal human predicaments” (Johnson-Laird and Oately 1992).^3

This helps us understand how the Rational Sentimentalists understand emotion.


Emotions, at least some set of them, are described here as being robust psychological kinds that result from evolutionary pressures. The emotions then are basic in the sense described in earlier chapters. They are evolutionarily derived sets of 'affective programs,' mental modules that accept a limited range of inputs and with stereotyped outputs. On this view that nearly all human beings share basic emotions with one another in much the same way we share other evolutionarily derived traits like bipedalism or noses. On this account emotions issue from discrete psychological mechanisms. This is an empirically testable claim.

Emotions might be differentiated from one another either by finding distinct physiological profiles for each emotion (presumably resulting from changes in a person's affective state) or perhaps by discovering distinct patterns of activation in the brain that correspond with each emotion. It is important to see why the Rational Sentimentalist makes this claim. Without this ‘distinctiveness’ thesis it would be hard to come to agree on a set of emotions that are universally shared. We would run into a similar problem as befell physiological theories of emotion in the second chapter. The Rational Sentimentalists avoid this problem by appealing to basic emotions. If each emotion has a distinct evolutionary history and issues from discrete physiological or neurological processes then we can overcome the problem. If basic emotions exist then we can bypass self reports and each emotion can be differentiated from one another on the basis of this kind of distinct marker.

---

4 One problem here, recall, was that physiological theories of emotion seek to find distinct patterns of physiological responses that could distinguish one emotion from another but that self reports become necessary to identify the relevant patterns of changes (if any). Since self reports are not obviously physiological this poses a problem for a purely physiological identification of emotions.
The Rational Sentimentalists believe their theory of emotion has a further advantage: it allows them to differentiate the basic emotions from the non-basic emotions. Non-basic emotions can be differentiated from the basic ones because they may lack one or more of these qualities: they may not be found across all cultures, they may not have distinctive physiological profiles or induce stereotyped changes in behavioral or motivational states, they may be dependent on certain beliefs not universally expressed, etc. On the Rational Sentimentalist view all non-natural emotions are derived from the natural emotions and are termed “sharpenings” of a natural emotion:

Sharpenings are constructed by specifying a subclass of instances of a sentiment in terms of some characteristic they happen to share. There are indefinitely many possible sharpenings, some of which already have names, such as homesickness and fear of flying. Cognitive sharpenings have a common thought or judgment, motivational sharpenings a common motive, and causal sharpenings a particular sort of elicitor.⁵

The paradigmatic list of basic emotions as construed by Rational Sentimentalists includes many familiar emotion terms: amusement, anger, contempt, disgust, envy, fear, guilt, jealousy, joy, pride, regret, shame, and sorrow. These emotions function as the building blocks out of which more complex or more specific emotions, 'sharpenings,' (and their corresponding evaluative properties) are constructed.

Although the Rational Sentimentalists argue that there are many evaluative concepts that appear susceptible to an analysis in terms of a constitutive relation

⁵ See D’Arms and Jacobson 2005. Rational Sentimentalists have made use of the concept of emotional sharpening in earlier research though they have since begun to back off from this conception. (personal communication).
between emotions and values. However it is up for debate, according to the Rational Sentimentalists, whether basic emotions constitute complex moral concepts like the good and the bad or right and wrong. Given all that has been said about Rational Sentimentalism and its conception of basic emotions (as a set of universal human emotions identifiable via characteristic physiological changes or patterns of behavior or motivational change) it might remain puzzling what special functional role basic emotions serve within an account of evaluative discourse. Why go through all the trouble of hitching oneself to a particular psychological and physiological theory of what emotions are in order to explain the way evaluative language works? What is the upside to committing themselves to the view that emotions, at least the basic ones, have distinctive physiological profiles? Before turning my attention to the empirical evidence on basic emotions, I wish to address these questions.

4.1 Rational Sentimentalism and the Role of Basic Emotions

I suggested in the previous chapter that basic emotions are especially useful to the Rational Sentimentalist due to the role that they play within their explanations of evaluative disagreement. I here expand on this explanation. Basic emotions have two unique properties that make them especially useful to the Rational Sentimentalist account of evaluative properties. Basically...we all have them and we are all stuck with them.6

Put more technically, each basic emotion, because it is the product of a shared

---

6 As I show in the final chapter, Peter Strawson makes an appeal to these two features of the 'reactive attitudes' to develop a Sentimentalist theory of moral responsibility.
evolutionary heritage, is universally accessible by all normal humans. Every human being has the ability to be afraid, ashamed, disgusted, jealous, joyous and the rest. One conclusion that can be drawn from this fact about the emotions is that all human beings are sensitive to the evaluative properties ascribed by each of the basic emotions. This gives the evaluative properties a degree of independence both from our own particular emotional responses as well as from the emotional responses found within any group or society. Because any person or society’s best articulation of the locus of concern of each emotion is contestable (to some degree), all human beings can be sensitive to the evaluative property of say, fear, even though different societies define fear (or the kinds of things that elicit fear) in ways that diverge. This makes the evaluative properties objective in one way while keeping the values essentially tied to human nature, and hence subjective, in another. This is a result of a commitment to the constitutive thesis. Reasons that speak to the fittingness of an emotion are also special; these reasons apply to all normal human beings but not to unemotional rational agents or agents with distinctly different kinds of emotional responses. An upshot of this conclusion is that disagreements about the nature of these evaluative properties, even radical disagreements, can be described as genuine disagreements about something both parties share. Recall the conversation between Frank and George in the previous chapter. The problem there was how best to explain the fact that Frank and George

---

7 Aberrations are bound to occur given the nature of genetic mutations and environmental effects. The final chapter will deal more squarely with emotionally aberrant individuals and how they have been used to develop theories of moral responsibility.

were engaged in a genuine disagreement about values, whether something counts as a good horror film, within a purely Sentimentalist framework.

I suggested then that if basic emotions are universal and can be identified from a distinctive set of physiological changes then, assuming both parties to the disagreement are normal human beings, we can verify empirically whether they are experiencing the same emotional response during the disagreement. This would be true despite Frank and George's potentially different conceptions of the locus of concern for fear or thrill. If the parties share the set of responses that comprise the emotional response at issue then their debate can focus on this shared response and then turn to a discussion of the best way to conceive of the evaluative property associated with the response. We can, quite literally, read the emotions off of their bodies (or their brains), point to them and say “look, you're both experiencing fear; you're both talking about things that are fearsome because you're both using the exact same physiological/neurological response as a guide.” Thus the appeal to basic emotions helps to secure a kind of objective status to the emotional responses and the resulting evaluative properties and this helps makes sense of radical disagreements as actual disagreements.

Lastly, if basic emotions are evolutionary derived reflexive modules and the evaluative properties they delineate are universally shared then they form an important class of human values. The Rational Sentimentalists can argue that we ought to take the evaluative properties more seriously because they are distinct from other kinds of
normative assessments and hence we ought to give more weight to the particular
normative dimension of human life that our emotions constitute. Thus one goal of the
account is to force us to take more seriously evaluative/normative concepts like the
shameful, the fearsome, the disgusting, the regretful, etc and to explain the wide
psychological and possibly moral roles that these evaluative concepts serve for us. If
the Rational Sentimentalist is right about the nature of basic emotions then it is clear
that the basic emotions would be useful for advancing these goals. They promise to
secure a common subject matter for our evaluative discourse due to the presence of
basic emotional responses in at least those values that are picked out by our emotions.
Unfortunately, much of the evidence in favor of the view of basic emotions
conceptualized by D’Arms and Jacobson has been undermined by recent and ongoing
research programs in psychology and the neurosciences. Basic emotions, if they exist,
do not manifest themselves as discrete changes in physiology or behavioral
expressions.

In the following sections I examine the evidence both in favor of and against
basic emotions. I conclude that the evidence for basic emotions is on shaky ground
though the claim is by no means defeated or shown to be false. It would be ideal if the
evidence for basic emotions were undeniable and robustly confirmed. This would
allow basic emotions to fill the roles described by the Rational Sentimentalists above.
Since this does not appear to be the case then a note of caution is in order.
Sentimentalists would do well to avoid committing their theories to controversial
views, like the existence of basic emotions, especially if the job that these emotions are being asked to fill can be realized by non-basic emotions. In the next chapter I argue that the roles that basic emotions have been asked to play can be satisfactorily filled by less robust conceptions of emotion. In particular that emotions as described by the Psychological Constructivist approach can fill these roles. Although the resulting Sentimentalism has a more limited, more subjective, scope than the Rational Sentimentalist's version this need not be problematic.

4.2 Basic Emotions: A Closer Look at the Thesis

Paul Ekman is best known for his research into facial expressions and for arguing that his research shows that certain expressions have evolved to express discrete emotions that are recognized by all human beings. This universal recognition is thought be a result of a shared evolutionary heritage. Ekman was strongly influenced by the theories of his mentor and adviser Silvan Tomkins and is perhaps the fiercest advocate in favor of a conception of basic emotions that most closely matches the Rational Sentimentalist's view of basic emotions. Ekman’s research, as I just noted, focuses primarily on facial expressions and the link between facial expressions and emotions. Ekman argues that:

The research on facial expressions has also shown the utility of conceiving of emotions as separate discrete states, such as fear, anger, and disgust, rather than simply as positive versus negative states or even more simply as differing only in respect to arousal.9

Ekman also argues that the research on facial expressions and emotion also seems to

---

9 See Paul Ekman “Facial Expression and Emotion” April 1993 American Psychologist Vol. 48, No. 4, 384-392
support the view that there are empirically verifiable discrete patterns of physiological changes. It is further thought that these discrete physiological patterns correspond with the discrete emotion terms we currently use. Lastly that these discrete patterns are attributable to a shared evolutionary history. Of course I have not yet said how or why facial expressions can prove so much. Ekman goes on to claim that:

Such a perspective would expect that emotion-specific changes in autonomic physiology would have evolved to serve the quite-different adaptations that are likely in emotions such as fear and anger.\(^{10}\)

Here again we see that the Rational Sentimentalist's conception of basic emotions bear a close affinity with the conception of basic emotions offered by Ekman. Discrete physiological profiles, partially caused by facial feedback—the movement of our faces as we contort it into smiles or frowns for example—can serve to pinpoint emotional states and discriminate one from the other. Furthermore, Ekman argues that basic emotions have evolved to serve particular functions and concludes that “[d]istinctive universal expressions have been identified for anger, fear, disgust, sadness and enjoyment” and that there is evidence to add contempt, surprise, and interest to this list.\(^{11}\) Ekman argues that each basic emotion can therefore be identified with specific neurological mechanisms called ‘affect programs’ and that each affect program is responsible for the recognition of emotionally relevant information. This information includes the production of physiological and behavioral changes unique to each emotion and also the characteristic facial expressions that accompany each basic

\(^{10}\) Ibid.
\(^{11}\) Ekman 1993.
emotion.

The single largest pool of evidence that is used to support the affect program theory relies on experiments designed to study the emotions that individuals associate with facial expressions. I want to give an overview of this experimental paradigm and then raise doubts about the conclusions drawn using it.

4.3 A Wink and A Smile

I do not intend to give a sweeping history of emotions scholarship in this section; instead, I want to more selectively focus on the more contemporary research that has been taken by many, including the Rational Sentimentalists, to support the kinds of conceptions of basic emotions that I have been discussing in this and the previous chapter. Specifically, as I mentioned above, I want to look at the work of Paul Ekman and his research on facial expressions. There are several distinctive claims that have gained widespread recognition. First is the proposed connection between facial expressions and emotions. Facial expressions are thought to be stereotyped and as evolved to signal specific emotions. Second is the existence of a distinctive set of physiological markers for at least some set of emotions. As I suggested in the previous section, Ekman’s affect program theory is often presented as the received view on emotions in psychological textbooks and so carries with it historical weight and influence. I focus primarily on his work and the research that has followed in its wake.

A standard facial expression study requires having subjects look at a photograph of a posed face. This picture is then presented along with a series of
emotion specific terms. Subjects are then typically asked to choose which emotion term best captures the emotion that is being expressed in the photograph. Each photograph focuses only on a person’s face. Subjects are denied any other information about the context in which the facial expression is made. The list of emotion terms that subjects are allowed to use in order to characterize the face is limited and is usually limited to the set of “basic emotions” terms. Subjects are only allowed to choose one emotion term (the one that “best describes” the facial expression) and they are then scored depending on whether they correctly identified the emotion term with the corresponding facial expression.

Most studies find high rates of agreement across different subjects and across multiple exposures. Rates of agreement on facial expression judgment studies typically vary between 70-90% of subjects agreeing to a particular facial expression-emotion pairing. For example, in studies where subjects are shown a prototypical ‘fear face’ and given a set of choices comprised by the standard list of basic emotions anywhere between 70-90% of subjects will converge on “fear” as the best description of the emotion that corresponds with the face that is presented to them.\textsuperscript{12} These high rates of convergence are found between Western (usually American) undergraduates. But the convergence does appear to generalize. When the same experimental paradigm is applied to non-Western literate subjects the rates drop somewhat but still remain high. When similar forced choice judgment studies are given to non-Western illiterate

\textsuperscript{12} It’s important to remember here that the “fear” expression corresponds with the experimenter’s description of the emotion the face is describing.
subjects from cultures relatively isolated from Western influence the rates of convergence drop even more so but still tend to remain above chance—sometimes significantly above chance.\textsuperscript{13}

The strength of this cross-cultural convergence between facial expression and emotional attribution provides evidence, it is argued, against the view that different cultures independently create facial expressions or their associated emotion concepts. In fact, such high rates of convergence among broad categories of subjects have been taken as evidence supporting the basic emotions view. These results are taken to imply, in accordance with the affect program view, that all human beings innately produce the same facial expressions in order to signal the same set of basic emotions and that the cross-cultural evidence further proves that all humans can recognize these facial expressions as markers for the basic emotions. Convergence, it is argued, cannot be accounted for without basic emotions. How else can we explain why isolated tribes in New Guinea associate the same facial expression with anger as American undergraduates?\textsuperscript{14} This facial expression experimental paradigm has been the most numerously replicated design in the psychological research on emotions and the convergence amongst subjects between facial expressions and emotion terms is as

\textsuperscript{13} In cross-cultural studies the emotion terms are translated into what are believed to be emotional analogues in the native language of study participants. Thus sad = triste (in Spanish and French). There have been questions about the semantic validity of this process (see Jean Harkins, Anna Wierzbicka's \textit{Emotions in Crosslinguistic Perspective} Mouton de Gruyter 2001). These issues will be addressed in the following section where I assess cross-cultural evidence.

\textsuperscript{14} The reference here is to Ekman and Friesen's “The repertoire of nonverbal behavior: Categories, origins, usage, and coding” \textit{Semiotica}, 1, 1969, 49–98 though see Pamela Naab and James Russell's “Judgments of emotion from spontaneous facial expressions of New Guineans” \textit{Emotion}, Vol 7(4), Nov 2007, 736-744 for new data and a critical analysis of previous research.
robust a finding as one can expect this kind of study to provide. For many years the cross-cultural convergence that resulted from the use of this paradigm lent credence to the basic emotions thesis and is still used to support this conclusion. Rational Sentimentalists like Justin D’Arms and Daniel Jacobson tie their account of evaluative properties to the empirical confirmation of basic emotions. For them it appears that Ekman’s research provides strong evidence in favor of their view. Ekman, and others agreeing with his research paradigm, provide the Rational Sentimentalist with evidence supporting the claim that a set of emotions is widely shared across all human beings and, to the extent that facial expressions are involved in an emotional experience, can help in distinguishing distinct behavioral and physiological changes that accompany each basic emotion. More recently however, the entire experimental protocol has been called into question and with it the basic emotions thesis.

### 4.4 Assessing the Evidence

James Russell provides several incisive arguments against the facial expression research program. I limit my attention to two distinct arguments made against the program and its conclusions about basic emotions. One objection focuses on the study design itself and casts doubt upon the internal validity of the results in general. Convergence, despite claims to the contrary, does not appear to show us that particular facial expressions are uniquely linked with a particular basic emotion. Russell comes to this conclusion after showing that subjects can and will converge on “wrong” answers when describing a standard facial expression. Furthermore Russell shows that
subjects will even fail to make the “correct” attribution when it is presented to subjects as an option. Both of these results stem from manipulations of the available choices subjects have in each study. Russell argues that Ekman and others, in effect, have stacked the deck in favor of their own conclusions by limiting a subject’s response options to a list comprised of basic emotions. One result of these conclusions, according to Russell, is that convergence in cross-cultural studies also appears to be artificially inflated by the experimental design.

4.5 Problems with the Forced Choice Paradigm

Russell wants to caution researchers (and philosophers) from taking too seriously the results of studies conducted according to the experimental design that I outlined above. Ekman and others have assumed that the high degree of convergence between subjects in these studies is best accounted for by assuming that the particular facial expressions under investigation are associated with the specific emotion that subjects tend to converge upon when asked to choose the emotion term that best describes the expression.

Convergence however is not a reliable indicator of the strength of the connection between a specific facial expression and the discrete emotion term used to characterize it. In order to show support for this claim Russell attempted to replicate

---

15 Correct and incorrect here means only that subjects either do or do not choose in accordance with predictions.
Ekman’s results by copying his experimental paradigm but also by varying two important parameters relating to the choices available to subjects.\footnote{Ekman’s experimental paradigm (along with the majority of facial expression judgment studies) is a within subjects forced choice judgment study. I will briefly discuss other facial expression studies that do not follow this model in the following section.}

In the first study Russell altered the set of forced choices that subjects had available to describe the emotion by removing the “correct” emotion term, according to Ekman, and replacing it with another basic emotion term. Why might this manipulation be important? On Ekman’s hypothesis each paradigmatic facial expression has evolved to express a discrete basic emotion; if this assumption is correct then subjects should be genuinely at a loss when choosing the most appropriate term if the “correct” emotion term is missing from their list of available options.\footnote{At times Ekman’s hypothesis seems rather stronger than this. He sometimes claims that facial expressions constitute emotions (at least in part) and that all emotions require facial expressions. Since this isn’t a claim shared by the Rational Sentimentalists (and also one that Ekman himself isn’t clearly committed to) I will ignore this added complication. See Ekman’s “Expression and the nature of emotion.” In K. Scherer & P. Ekman (eds). Approaches to Emotion. Hillsdale, NJ: Erlbaum, 1984 pp. 319-344. for the stronger view but see also his “Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique” Psychological Bulletin 115(2) 1994:268-87 for an apparent denial of this claim.} One might expect that subjects would tend to choose rather haphazardly between the corresponding negative or positive choices (depending on the original face) when the “correct” answer is missing. For example, subjects that are presented with a face deemed to represent “anger” might then have to choose from the following list of emotions to describe the photograph: happiness, surprise, contempt, fear, and interest.\footnote{Note that all of these options are ‘basic’ emotions according to Ekman and the Rational Sentimentalists.} When these options are coupled with Ekman’s hypothesis about facial expressions, a hypothesis shared by the Rational Sentimentalists, that each emotion
has evolved to serve a discrete purpose and has discrete markers we can better point out the problem. Each emotion, recall, is modular. Mental modules are very particular about what they accept as inputs—they are domain specific—and with what they return as outputs. They are separate from, and not influenced by, other psychological processes. They are what philosophers and psychologists call 'informationally encapsulated.' Given the characterization of emotions as affect programs and of affect programs as mental modules one would expect subjects to be at a loss to provide a ‘best’ answer when presented with Russell's question about facial expressions. Happiness and interest might be rationally ruled out because they are clearly incorrect (being positive emotional states) but we can expect a rough distribution amongst surprise, contempt, and fear given the options (because all these emotional responses are negative in some respect). This is not, however, what Russell found. Instead, Russell's results showed that subjects consistently converged on a particular emotion term with convergence rates as high—and in some cases even higher—than the rates of convergence that Ekman used to derive his set of basic emotions.

For each facial expression, subjects in Russell's study managed to converge on an emotion term to describe the facial expression. These convergence rates are quite robust and ranged from 76.2% - 96.3%. If we maintain the hypothesis that a facial expression is an involuntary, evolutionarily derived, signal that expresses a discrete emotion that subjects can identify instinctively and then report on, then these results

---

21 Russell and Fehr 1994 (4)
22 Ibid.
may lead us to draw some disturbing conclusions. Facial expressions hypothesized to clearly signal one particular emotion seem to signal, with equally high rates of subject convergence, a very different basic emotion. This is the case if we take convergence upon an emotion term as an indicator of the actual emotion being expressed by a facial expression. One lesson to draw from Russell’s study is to exercise caution when interpreting the results of forced choice response studies that purport to find links between particular facial expressions and particular emotions. It seems that the rate at which subjects converge on a response in a forced choice study is either not a good indicator for the emotion expressed by a facial expression—hence the convergence on different emotions for the same facial expression—or facial expressions may express multiple emotional states at the same time. Either thesis undermines Ekman’s claim to have found basic emotions via cross-cultural convergence. Indeed Russell’s study shows that the data showing intra-cultural convergence needs to be re-examined.

Russell’s first study undermines the temptation to take evidence of convergence on a set of emotion terms as evidence that this set of emotions is basic. His second study warns us against thinking that facial expressions are a good indicator for discrete emotion categories at all. Russell argues that these studies should lead us to think that faces may express various kinds of information. Although some of the information that facial expressions provide us is affective not all of it appears to be. This seems true even for the most stereotyped facial expressions used in judgment.

23 As I’ll argue later, if facial expressions do not signal particular emotions but instead provide us with more general information about a subject’s affective state (such as their relative position within the space of ‘core affect’) then this data need not disturb us.
studies.

In a second study Russell modified the options in the forced choice study to include non-emotional but affective states with a similar valence charge as the 'correct' emotion for a facial expression. Included among each subject’s response options was the presumed “correct” emotion term. For example, an Ekman “anger” face was presented to subjects who were then allowed to choose from the following list of affective descriptors: frustration, determination, anger, hostility, hatred, jealousy, and pain. Under this condition it would appear that a basic emotions theorist who agreed with Paul Ekman or Justin D’Arms and Daniel Jacobson ought to claim that subjects should overwhelming prefer “anger” as the correct response when presented with an “angry” facial expression. The point again returns to the supposed modularity of affect programs. If affect programs truly are modular in the sense described earlier then an angry face just is the perfect domain specific input for the anger affect program. If affect programs, including their outputs, really are informationally encapsulated then top down effects like a subject's beliefs about her emotions or her expectations about the demands of the study should not be able to influence her judgments about facial expressions. Just as we cannot talk ourselves out of seeing the two lines in the Müller-Lyer illusion as being of different lengths—even when we know they are the same length—a subject should automatically respond to the anger face as an anger face.

In Ekman’s studies it was the convergence on “anger” by his subjects that helped him to confirm that the facial expression presented to subjects did indeed
express that particular emotion. Once again however Russell’s results contradict the expected conclusion that would result if Ekman's theory of affect programs were true. Russell's data helps undermine the claim that convergence can be taken as evidence that the emotions being converged upon are widely shared. Furthermore they undermine the claim that facial expressions signal specific emotions as a result of evolutionary processes.24

Russell's results here are interesting. For example, subjects shown two different photographs of an angry face, A and B, fail to choose “anger” as the best description of the facial expression even though both A and B were photos from Ekman’s archive of prototype anger faces. This is puzzling given the strong facial expression-emotion link posited by Ekman. If competent English speakers fail to converge on anger as the best description of an “angry” face then it seems very hard to validate the results gained earlier by Ekman and his colleagues that were derived from forced choice studies. It is even more risky to generalize those results across all cultures and time periods. In Russell's second study subjects seemed to find the face presented to them to express distinctly different kinds of states some of which were emotional, like frustration and anger, and some that were not, like determination.

One easily dismissed reply to Russell’s data would be to claim that convergence would remain robust using a different experimental design. Facial expression studies have been conducted without the use of the forced choice response format and the results here are less favorable for the affect program view. Many of

24 Russell 1993 (8)
these studies allow subjects to write in their own best responses to each facial expression. The results in these “free choice” experiments, in which subjects are allowed to choose their own terms to describe facial expressions, are much more difficult to compile and the results are far from promising for the basic emotions theorist. For example, it is very difficult to group together disparate answers under a common rubric in order to make the claim that subjects labeled a facial expression with the same emotion. If one subject writes ‘a person on a rainy day’ and another writes ‘crying’ for the same facial expression, are both subjects claiming that the face expresses sadness? In large part this depends on how the data is interpreted. In order to achieve the high rates of convergence found in the forced choice format, researchers must group together disparate (and sometimes dubious) responses as being the same. Furthermore the heterogeneous nature of subject responses should by themselves be a cause for concern for the affect program view. Given that the lion’s share of Ekman’s basic emotion evidence is derived from the forced choice response format, Russell’s data strongly undermines the conclusions drawn by Ekman and others and casts strong doubt on the view that basic emotions, should they exist, are best understood as modular affect programs.

4.6 Study Design Issues Revisited

Recall that the largest single source of evidence for the basic emotions thesis

---


26. The single largest pool of data conducted in the psychology of emotions is comprised of facial expression recognition studies of just the sort I’ve been describing.
comes in the form of an appeal to the robust nature of convergence in facial expression judgment studies. Basic emotions are thought to be the best explanation for these high rates of convergence. A different explanation, and one that I have been attempting to detail here, is that these high rates of convergence could be artifacts of the judgment studies themselves. The forced choice nature of many of these studies casts doubt on the generalizability of the results. Much like a magician can selectively “force” a participant to choose a particular card from a deck when told to ‘pick a card, any card,’ the forced choice design can funnel convergence onto a range of emotions depending on the options that subjects are provided. Heterogeneous responses also undermine the modular nature of affect programs. Aside from the forced choice nature of the typical judgment study, there are other factors at play that are likely inflating the recognition scores.

In this section I want to point out two other troubling features of the standard judgment study. Judgment studies with the highest rates of convergence make use of a within subject study design. In a within subject design, each subject is exposed to the entire set of facial expression photographs and is later asked to judge each one separately. The nature of this design tends to increase the rate of convergence for the simple reason that exposure to a variety of different photographs of human faces leads subjects to focus on the differences between each face. This may make subjects reluctant to label two faces as ‘angry’ for example, because of the perceived demand of the task itself instead of the subjects’ perception of the emotion being expressed. On
the other hand, between subject studies are one-off deals. With this design a subject is shown only one facial expression and asked to judge it without being exposed to the entire set of facial expression photographs. There is some indication that rates of convergence differ—by as much as 20%—depending on whether a judgment study makes use of a within subject or between subject study design. Also, some emotions are much less likely to be chosen as responses depending on study design. Subjects are more likely to say that a face expresses contempt, for example, when they have seen many other instances of other facial expressions (a within subject design). On the other hand, if presented in the context of a between subject format ‘contempt’ faces are most often described as ‘angry’ faces. As in the critique of the forced choice response format, a within subject design by itself should not be taken to invalidate the facial expression data (far from it) but we should be growing ever more cautious about what kind of theory of emotion these kinds of studies should be seen as supporting. If different elements of the design itself are responsible for artificially raising the rate of convergence in the study itself then the evidence for universality is far from secure.

4.7 Emotions in the Face

A body of evidence is amassing that questions the ecological validity of facial expression judgment studies. Two features in particular have been highlighted as problematic: the use of posed (rather than spontaneously generated) facial expressions and the lack of consideration for the effect of context on our understanding of the link

---

between facial expressions and emotions. Addressing these two issues, some claim, would show us that even highly stereotyped (universal) facial expressions do not occur only in specific contexts and do not express the same emotions in all contexts.

The facial expression studies I have been discussing in this chapter are well known and Ekman’s research in particular is familiar even to non-academic audiences. In particular, the facial expressions that Ekman has come to study and categorize, along with the details of his research, are taught in many introductory psychology courses and in some high school contexts. As I have already discussed, because Ekman’s theory is considered the received view, university students across the world are likely to be familiar with (and influenced by) this view of emotions and the meaning of facial expression.

The typical Ekman facial expression photograph consists of a tightly cropped photograph of a person’s face against a neutral monochrome background. Each participant whose face appears in these photographs is not simply instructed to ‘make a happy face,’ for example, but instead directed to move particular facial muscles in very specific arrangements in order to produce what Ekman and colleagues argue the ‘happy face,’ the evolutionarily derived signal for the internal emotional state of happiness, consists of.

In short, the photographs that subjects are exposed to are sets of highly specific posed human faces. Although it makes some sense to desire stereotyped stimuli in scientific experiments, we may wonder why subjects would require the degree of
coaching that is given in order to produce a facial expression Ekman and colleagues themselves argue is innate, universal, and the product of a mental module. Part of the reasoning behind posing facial expressions is that it is very difficult to elicit genuine facial expressions in a laboratory setting and so posed facial expressions are presumed to be the next best thing. This line of reasoning rests on an interesting assumption however. It assumes that a posed facial expression is communicatively equivalent to the facial expression subjects would spontaneously make while actually feeling an emotion. For example, it assumes that spontaneously generated happy faces share the same features as the posed happy face, that all sad faces share some set of features, and so on. Is this true?

Luckily some facial expression judgment studies have attempted to use spontaneously produced facial expressions. The results of these studies closely match the conclusions that resulted from shifting from a forced choice to a ‘free response’ format. The use of rigidly cropped and stereotyped photographs, that is, the structure of the experiment itself, seems to be leading the experimental outcome toward response convergence. Psychologist Carney Landis conducted one of the earliest, and most bizarre, studies on spontaneously produced facial expression in 1924.28 The aim of Landis' experiment was to test the degree of similarity between different subjects’ spontaneously generated facial expressions. To have subjects generate spontaneous facial expression in a laboratory Landis devised several different situations meant to

---

elicit particular emotions. The hope was to discover whether subjects experiencing the same emotion would produce the same facial expression as other subjects put into the same emotion eliciting situation. Subjects were instructed to perform different tasks under Landis’ explicit direction. While some of the tasks were relatively straightforward and non-controversial—for example, subjects were instructed to tell a lie, smell ammonia and so on—the stimuli became progressively more emotional and more coercive. Fireworks were ignited behind the subjects in order to elicit fear; pornographic photographs were also shown to subjects in order to produce sexual arousal. In another instance subjects were instructed to reach into a covered bucket with live frogs inside to elicit disgust. The most bizarre task involved Landis handing subjects a live white lab rat and a knife. Landis then instructed each subject to decapitate the lab rat. Although many subjects initially protested, Landis continuously repeated the command, eventually 75% of subjects complied and decapitated the lab rat though some did so very reluctantly. Most interesting for my purposes is the fact that Landis made sure that each subject had a series of photographs taken of their facial expressions throughout the course of each task including as they decapitated the rat. Although most subjects reported feeling sad and frightened as they killed the rat, they displayed a dynamic range of facial expressions as they actually performed the decapitation. Some subjects laughed and smiled, others contorted their faces into patterns not recognizable as a prototypical facial expression for any emotion and some

---

29 Although this study predates the infamous Milgram experiment, the results (in terms of rates of compliance) are surprisingly similar. I should note, Landis personally decapitated the remaining 25% of the rats in front of each subject that refused his instructions.
failed to express, via a facial expression, any emotion at all. Landis found little
evidence to suggest that particular facial expressions were used to express particular
emotions.

Putting aside the more sensational aspects of Landis’s experiment, the study
itself is useful because it represents an early attempt to test an important assumption
driving most modern facial expression judgment studies. Few studies I have
mentioned have investigated the question of whether subjects undergoing a specific
emotional experience produce a stereotyped facial expression. Most judgment studies
rely on this kind of implicit assumption: that happy people smile, that sad people
frown and so on. The Landis study however suggests that this assumption may not be
so safe for us to make. It may be that the assumption does not reflect our actual
emotional experience. Much like the All Components Necessary (ACN) theories of
emotion failed to do justice to our actual lived experience with emotion, it may turn
out that the stereotyped expressions of sadness, happiness, fear, anger and so on
function better as exemplars or prototypes of emotions rather than being the way in
which many, or even most, of our emotions are expressed—even via the face.

In particular there are two assumptions about the link between facial
expression and emotion that need to be verified before we should feel comfortable
taking judgment studies as support for the basic emotions view. The first assumption
has already been put on the table: that subjects undergoing an emotional experience
reliably express that emotion by producing a stereotyped facial expression. Secondly,
what would happen if we were to take Landis’ photographs and ask another set of subjects to judge the emotion that each of Landis' subjects was experiencing as they decapitated the rat? Would they be able to judge, from the subject's face alone, what emotion each of them reported that they were feeling? This question leads to the second assumption: that we can easily recognize the emotion someone else is feeling by looking at each subject’s *spontaneously* produced facial expressions. Both of these assumptions are open to doubt. If either is false then we have more reason to doubt that emotions should be understood as informationally encapsulated domain specific affective modules. In fact, recent evidence suggests that both assumptions are likely to be false.

In a pair of studies Jose Fernandez-Dols and Maria Ruiz-Belda probed each of these assumptions. Fernandez-Dols observed subjects under conditions likely to elicit one particular emotional response very strongly: happiness. In one study he observed gold medal winners at the Olympic ceremonies prior to, during, and then after they received their medal on the Olympic podium. Another study involved naturalistic observation of a group of soccer fans as they watched an important match between their home team and a rival. In the first instance it is clear that winning a gold medal at the Olympics games is likely to elicit bouts of happiness in the winner. It seems a safe assumption that for most of us winning an Olympic gold medal would represent the achievement of what is likely a lifelong dream. In the second study each subject’s facial expressions were studied in response to positive and negative events during the
game (the home team scoring a goal for example). Subjects were watched over the
course of several matches and verbal reports were taken to indicate each subject's
emotional state during these important moments. In this case, emotional events were
further subdivided depending on whether there was clear social interaction during the
event or whether the subject was emoting privately. Social interaction, in this
context, meant whether the facial expression was intended, by the subject, to be seen
by others or whether it was made spontaneously without thought to a social context.

In both studies an interesting fact emerged; the rate of smiling expressions
produced by each subject varied depending on social context. Prior to having the
medal placed on their necks, while waiting behind a podium, gold medal winners
smiled less than 5% of the total time. However, when all eyes are on the gold medal
winner, at the moment when they are having the medal placed on their necks, subjects
were ten times more likely to smile. Immediately following this moment however, the
rate of smiling drops to less than 2%. Presumably the winners were not happy only at
the moment of having the medal placed around their necks but throughout the entire
episode. One could imagine that the happiness would last far beyond the end of the
ceremony in fact. If so, then the fact that medal winners only produced a smiling face
during some moments of their happiness stands in need of explanation. In the soccer
study, rates of smiling varied according to social context. During interactive episodes
of happiness subjects smiled nearly 50% of the time. When an episode of happiness

See Jose Fernández-Dols and Maria Angeles Ruiz-Belda’s “Are smiles a sign of happiness? Gold
medal winners at the Olympic Games.” Journal of Personality Social Psychology 69:1113-19 and
Maria Angeles Ruiz Belda et al.’s “Spontaneous facial expressions of happy bowlers and soccer
fans” 2003 Cognition and Emotion.
was not interactive, the rate of smiling dropped to less than 10% even though in both cases subjects reported being happy.

What these three studies suggest is that the link between particular emotional episodes and particular facial expression may be weaker than is commonly assumed by those basic emotions theorists who rely heavily on expression based studies to support their claims. Landis’ study suggests that even strong negative emotions may elicit bouts of spontaneous smiling while the pair of studies by Fernandez-Dols and Ruiz-Belda indicates that happy subjects do not smile spontaneously with any regularity unless they are in a social situation where it is important to explicitly make others aware of the fact that you are happy. It is important for happy soccer fans to let one another know they are happy but recall that extremely happy soccer fans produced stereotyped smiles only 10% of the time when they were not in a social context.

Spontaneous smiling, it seems, is not reliably produced by happiness and even strong negative emotions can produce what looks like spontaneous smiling. Although it is premature to generalize the results of these studies to all emotions and all emotional expressions, they do force us to examine common assumptions about the link between emotions and facial expressions.

If a strong link does not exist between particular facial expressions and particular emotions then one important conclusion that can be drawn is that the convergence that arises from facial expression judgment studies is potentially an artifact of the study design itself and not a robust truth about the kinds of emotions
that human beings have. What I now want to argue is that cross-cultural data showing convergence inherits this problem and it sufficiently undermines the facial expression data to force philosophers to look elsewhere for evidence in favor of basic emotions.

4.8 Cultural Contact

It may seem as if I am belaboring the point about facial expressions. In the last sections of this chapter I consider in more detail the physiological and neurological data that has also been thought to support the basic emotions view. It is important however to challenge the facial expression findings because they have historically been seen as a strong source of support for the universality claim. In particular the myriad of cross-cultural facial expression judgment studies conducted over the last thirty years have appeared to support the claim that subjects around the world can recognize that specific facial arrangements express specific emotions. The format in these studies is largely similar to the intra-cultural studies that I have already been describing. Cross-cultural studies have demonstrated that rates of cross-cultural convergence for pairing an emotion with a facial expression occur at rates above chance. As I mentioned early on this chapter, rates of convergence are sometimes dramatically above chance. The question on the table at the moment is how to interpret this finding.

Some have claimed that the best explanation for these rates of convergence is that basic emotions exist, that basic emotions are innately linked to specific facial expressions, and that we are able to quickly and easily interpret someone’s emotional
states by looking their facial expressions. This, I think, is too hasty a conclusion to draw from the cross-cultural data.

At first blush the data do seem convincing. Russell argues that among literate subjects in western cultures the highest rates of convergence approach 100% for emotions like happiness. Even the emotion terms with the lowest rate of convergence are near 80% for states like fear. Among members of literate non-western cultures, most subjects' rates of convergence are still surprisingly high. Happiness again is by far the most agreed upon emotion with around 90% of subjects converging on a smiling face as representing happiness.\(^{31}\) Anger, the lowest agreed upon emotion in non-western literate subjects, still manages to have more than 60% of subjects across cultures converge on it as the emotion being expressed. Even among isolated illiterate non-western peoples, what would seem to be the best test case for basic emotions, the rates of convergence are impressively high. Happiness rates above 90% convergence while surprise, the lowest agreed upon emotion in this group, approaches 40% convergence. In the case of surprise, the 40% convergence is approaching but still above chance guessing. In all other cases, emotional recognition rates drop for this final group but, like surprise, the overall rates remain above what we would expect with random guessing.\(^ {32}\)

\(^{31}\) Much of these studies were conducted in each subjects native language so the claim here must be attenuated to something like: 90% of subjects converge on a translation for the English word 'happy' when presented with a smiling face.

I have already outlined a few reasons why we should be careful about interpreting rates of convergence in the context of forced choice judgment studies. Those worries remain in place when we extend the forced choice paradigm to studies across cultures. The most interesting groups in these studies, and the groups with the lowest convergence rates, are the groups composed of isolated illiterate non-western peoples. In the case of the other two groups the test subjects are overwhelming composed (naturally) of undergraduate college students living in their respective countries. This population is far more likely to be aware of facial expression judgment studies (in particular) and are more likely to have greater exposure to Western cultural influences (television programs, movies, magazines, music, and so on). This kind of exposure would function as a powerful primer to the facial expression and emotion pairings being tested. If culture does play a role in how we understand emotions and the link between emotion and facial expression then cultural contamination may potentially mask these results. The only population where this is not true, isolated illiterate non-western peoples, is also the population with a significantly lower rate of agreement among facial expressions. Although convergence in this group remains above chance this does not necessarily imply the existence of basic emotions. A loose link between facial expression and core affect, for

---

33 Russell 1995
34 They are also the group which calls for the most care when interpreting results. The basic emotion theorist could argue, for example, that lack of convergence in these cultures is due at least as much to problems understanding the task or problems with translation than it does to problems with the basic emotions thesis. Although a live option this claim is undermined when accompanied by the a presentation of the problems with the paradigm itself along with the lack of non-judgment study emotion specific physiology. I turn to that question in the next section.
example, would be enough to differentiate many emotion categories from one another without the need for basic emotions. Positive affect, for example, while not identical with our conception of happiness—which includes prototypical elicitors, behavioral cues, and expressions—would be enough to differentiate the positive happy face from the other negatively valenced faces. This would make it easier to converge on happiness as an indicator for smiles than as indicators of other emotions. There is, for example, a nearly 30% drop in rates of convergence between happiness and the next highest emotion on Ekman's list.

Although it is unlikely that cross-cultural data functions as the lynchpin of support for basic emotions that it was once believed to, it does need to be taken seriously. Even if we control for issues of experimental design—the issues brought to light in this chapter—it is unlikely that rates of convergence for all emotions would fall to chance levels. This would imply that a fully constructivist theory of emotions is false. Cultures do not create emotional concepts ‘out of nothing’ as it were. If they did then we would expect far less convergence than the data bear out. Something must explain why, even in the absence of discrete basic emotions, cross-cultural rates of convergence remain above chance.

I pause here to note that Psychological Constructivism, the theory outlined in the second chapter, is not committed a fully constructivist theory of emotion. As we saw in chapter two, the essence of Psychological Constructivism is not to deny the existence of universal components of human emotion like core affect. The view only
denies that the same set of emotional concepts is universal. This is compatible with the claim that the core affective system that grounds emotional concepts is universal. If core affect is in a loose sense linked not to the whole facial expressions but instead to more specific facial muscle movements, as advocates of a componential analysis of facial expressions suggest, then that could explain why the judgment studies do not show chance level convergence.35

My aim at this point has been to cast doubt on the link between facial expression judgment studies and basic emotions. I have not provided the final nail in the coffin for basic emotions but I hope to have lead us to move away from using facial expressions studies as the main pillar of support. If the face does not tell us whether emotions are basic, then what might? We have already seen one response in the second chapter. William James’ theory of emotions required the existence of a set of emotion specific physiological changes that served to differentiate one emotion from the others. Emotion specific physiology has sparked great interest among emotions researchers for the same reason. If stereotyped universal physiological or neurological changes could be identified that correspond to different emotions then it really would seem that a set of basic emotions would be the best explanation for this fact. Unfortunately for the basic emotions theorist, there is reason to doubt that this kind of highly specific emotional physiological changes will be found.

4.9 Other Clues: Physiology and Neuroscience

If the results of facial expression studies do not cast definitive support in favor of the basic emotion claim then it is possible, some have thought, that other avenues might yield more favorable evidence in the thesis’ favor. In the second chapter I explained and critiqued what I then called ‘physiological theories’ of emotion. I argued that that these views were committed to the there being a set of discrete physiological profiles (DPP) that could be used to individuate each emotion. Because this family of theories identifies an emotion with the physiological changes that occur while we experience it, it must be the case that the physiological changes are stereotyped and unique to each emotion. If this were not true, I argued, then we would not have any way of knowing what emotional states we were actually in by looking only at a subject's physiology. This would be true because a physiological profile might be consistent with many different emotions and so we would need to turn to further evidence, self reports for example, in order to differentiate emotions from one another. Physiological theories of emotions are not necessarily committed to basic emotions, it could turn out that physiological profiles vary from culture to culture for example, but they tend to take on basic emotions as a commitment. It is often thought that the best explanation for stereotyped physiological profiles for each emotion, if any were to be found, would be best explained by those profiles being basic. The relationship works in the other direction also. Many basic emotion theorists also commit themselves to the claim that emotion specific physiological profiles exist.
Justin D’Arms and Daniel Jacobson, as I have already noted, explicitly rely on the existence of emotion specific physiology in order to support their arguments regarding the role of emotion in radical evaluative disagreement. Recall their solution to the problem of evaluative disagreement. Two parties, according to the Rational Sentimentalists, can disagree about both the extension of an evaluative property—what kinds of things are funny—and also about the nature of the evaluative property itself—what it means for something to be funny—but can still engage in a meaningful evaluative dispute. This was because the Rational Sentimentalists argue that the two parties share a basic emotional response that helps to constitute the evaluative property. That two parties shared a basic emotional response was independently verifiable because each basic emotion was differentiated at least in part by emotion specific physiology. If some emotions have universal and automatic physiological profiles then that is a good reason, it would be thought, to believe that those emotions are basic.

The brain too has been examined in the hopes of finding discrete universal patterns of neurological activation that might be useful in discerning one emotion from another. My main task here will be to note that, although the data is far from fully in on this issue, recent meta-analyses of the studies in this field do not support the claim that there are discrete physiological profiles for emotion. This is true for both purely physiological or neurological profiles. Interestingly, much of the data seems amenable to a dimensional as opposed to a basic emotions analysis and hence can be equally
well explained without reference to basic emotions. On a dimensional analysis emotions are not seen as necessarily discrete but as ranging along a continuum. Core affect represents one kind of dimensional analysis though not of emotion but of the more basic affective element of emotion. Dimensional analyses of affect appear better poised to explain the results of meta-analyses also.

In fairness to the discrete physiological profile claim, early evidence for DPP seemed strong. In the last 50 years many have claimed to find sets of physiological markers that differentiate one emotion from another. These markers have ranged from patterns of activation of various facial muscles, changes in heart rate, blood pressure, finger temperature, eyeblink rate, pulse, respiration rate and depth, and many others. Although challenges to each of these individual studies have been made, a meta-analysis of all of the relevant studies had, until somewhat recently, been lacking. Two major meta-analyses, both published within the last ten years, have helped to address the question of whether emotion specific physiological and neurological profiles are suggested by these studies and both meta-analyses suggest, at the very least, that those profiles are not identifiable given current technologies.

John Cacioppo and his colleagues conducted one of the studies that has helped to fill this gap. While not entirely negative, his group's results suggest that nearly all of the supposed emotion specific parameters failed to show a statistically significant

---

connection with discrete emotions. Cacioppo's group focused exclusively on studies conducted within the last twenty years that claimed to have discovered at least one emotion specific physiological profile. Data for all these studies were compiled in a search for any statistically significant correlations between emotion and physiological markers that would rise from the mountain of research conducted in the last two decades. Meta-analyses are in many ways always at risk of being too messy. It is hard to account for the fact that different studies use different methods of analyzing variance and significance; different studies may operationalize variables in importantly different ways; and they all may interpret their results differently when they look for significant correlations. Despite all of these potential problems meta-analyses can be useful tools for getting an overall view of the research being conducted within a field. This is exactly what Cacioppo and his colleagues sought to do. The most robust physiological changes found, and the ones most predictive of the kind of affective state a subject was in, were physiological changes associated with positive and negative emotions in general and not with particular ‘basic’ emotions:

In sum, the meta-analyses indicated that even a limited set of discrete emotions such as happy, sad, anger, and disgust cannot be fully differentiated by visceral activity alone, but follow-up meta-analyses did suggest that the negative emotions in this literature are associated with stronger ANS responses than are the positive emotions. Thus the evidence for the visceral differentiation of emotion, like that for incipient facial differentiation, is clearer when positive and negative emotions are contrasted than when discrete emotions are contrasted.

---

37 Cacioppo et al 2000 and also Larsen et al 2008
38 The more recent Larsen et al. meta-analysis is more inclusive than the earlier Cacioppo 2000 analysis.
39 Cacioppo et al. 2000 (184)
As indicated by Cacioppo, the search for basic emotions grounded in physiological changes parallels the search for emotion specific facial expressions. Large scale and significant differences exist between positive and negative emotional states and these differences can be tracked and identified by observers via facial expressions. Positive and negative emotional states are also easy to track for the individuals in the grip of these emotional episode by monitoring their physiological and phenomenological states. This kind of positive versus negative distinction can explain why, for example, subjects in Russell’s study failed to associate the “angry” face with happiness and why Cacioppo did find clear differences between positive and negative emotions but failed to find emotion specific differences.  

A more recent meta-analysis conducted by Jeff Larsen and his colleagues, which included Cacioppo, echoes Cacioppo's earlier results. Both meta-analyses fail to vindicate empirically the existence of a set of emotions that are universally shared and universally recognizable, at least by way of discrete physiological profiles.

Outside of physiological measures, the brain itself has been thought to harbor emotion specific patterns of activation. If true then these patterns could serve the functional role of differentiating between emotional states and once again validate the basic emotions claim. Compiling fifty five neurological studies of emotion, K. Luan

---

40 Core affect was defined as a two dimensional affect space comprised of valence and arousal. It is not surprising that the large scale difference between a positive and a negative emotion would result in detectable physiological changes. It is unclear if physiological differences were associated with emotions laying on opposite ends of the other dimension of core affect: arousal. It seems likely that high arousal emotions would be detectably different form low arousal emotions but no explicit report of such differences can be gathered from these studies.

41 See footnote 110 above.
Phan and colleagues conducted a meta-analysis of neurological studies that investigated the question of emotion specific patterns of neural activation. Phan did not find specific neurological nuclei associated with discrete emotional responses.\(^\text{42}\)

Interestingly, even brain areas like the amygdala, long though to be associated with fear, were found to respond to emotional stimuli of opposing valence:

Amygdala activations occur throughout various evocative stimuli, including fear faces (Morris et al., 1996; Breiter et al., 1996; Phillips et al., 1997), aversive pictures (Irwin et al., 1996; Taylor et al., 1998; Simpson et al., 2000), as well as sad (Blair et al., 1999) and happy faces (Breiter et al., 1996), and positive pictures (Hamann et al., 1999). A positive correlation of blood flow in the amygdala was found with subsequent recall of pleasant pictures (Hamann et al., 1999): in that study, the amygdala activated to both pleasant and unpleasant pictures. Thus, the amygdala may not exclusively respond to affectively laden stimuli, but may respond to meaningful stimuli in general.\(^\text{43}\)

Although some brain regions were found to be more likely to activate during an emotional response, Phan’s meta-analysis failed to find any specific patterns of brain activation that function to differentiate one discrete emotion from another. These results have also been confirmed by a more recent meta-analysis of neuroimaging studies conducted by Lisa Barrett and Tor Wager.\(^\text{44}\) Results like this have led even


\(^{43}\) See Phan et al 2002 (339). As always, a note of caution is in order. More precise imagining technologies may associate even more specific regions within the amygdala that are only activated by fearful stimuli. Phan himself notes this possibility (343). The same holds true for all neuroimaging data however and as of yet, our best evidence points away from emotion specific patterns of activation.

\(^{44}\) See the previously mentioned work by Lisa Barrett and Tor Wager “The Structure of Emotion: Evidence From Neuroimaging Studies” *Current Directions in Psychological Science*, 15, 2006 p. 79-83
formerly strong supporters of basic emotions to soften their views. Jack Panksepp, for example, has recently articulated his theory of emotions in the following way:

Many aspects of emotion emerge from genetically engrained brain processes that are homologous in all mammals. Other aspects emerge from the epigenetic interaction of those processes with ecological and social environments and yet others from our human ability to semantically conceptualize issues of importance to us.⁴⁵

Indeed he goes on to argue for claims in rough agreement with the theory of emotion I promoted in the second chapter and whose implications I explore in the next:

[I]t is likely that affective concepts such as “guilt,” “shame,” and “sympathy” do not really exist as fundamental processes, even in the human brain, but they can easily arise as derivative processes based on social learning that weaves such basic feelings as separation distress and social bonding into more complex sociocultural realities. Once evolution endowed our ancestors with basic “integrative systems for social affect” (Panksepp, 1998a), along with a certain amount of complex, multimodal association cortex, the emergence of language could construct many concepts from the bipolar dimension of feelings engendered by social isolation and social warmth/solidarity.⁴⁶

As a philosopher I feel only half prepared to digest and interpret scientific data in the way I have been doing throughout this chapter. When philosophical theory makes contact with the world of empirical research the best one can do is survey the literature

---

⁴⁵ See Jaack Panksepp's 1998 *Affective Neuroscience: The Foundations of Human and Animal Emotions* (Series in Affective Science). Oxford University Press, New York, New York (137) for this discussion though see his “Affective consciousness: Core emotional feeling in animals and humans” *Consciousness and Cognition* 2005 for what appears to be a more robust defense of basic emotions. Panksepp differentiates between various levels upon which we could theorize about emotions. There are 'basic' neurological systems of affect that most mammals share and which are given labels that sound like emotion terms (lust, fear, panic, rage, play, etc) but this is deceiving. The level upon which much of my own discussion has taken place is at the conceptual level: what do our emotion terms mean? Fear, at the level which Panksepp considers basic, is devoid of the conceptual content we are accustomed to giving it. At this most basic level it seems fair to suggest that Panksepp's labels are picking out points within the space of core affect, a region where content is at a minimum, than the more conceptually rich emotional ascriptions under discussion now.

⁴⁶ Panksepp 1998 (138)
by casting one’s net out and getting a sense for the state of the field. I have cast my net and this chapter represents my catch. The state of the field is very much in the middle of an active debate about the meaning of its core concepts—emotion, affect, arousal, basic—and so a note of reserve should be added to any empirical claims about the nature of the existence of something like basic emotions. What I am sure of is that evidence for affect programs is weaker than once presumed. Affect programs, understood as modular informationally encapsulated programs composed of discrete neurological, physiological, and behavioral patterns, including facial expressions, do not appear well supported by current evidence. There is much reason to think that data that was taken to support this view, like the facial expression studies, is compatible with less robust theories of emotion. Since the largest pool of supporting evidence stemmed from universal behavioral patterns (facial expression studies) and the ecological validity of that evidence has been undermined, it remains to be seen whether neurological or physiological emotional profiles will emerge. Evidence to date suggests otherwise.

Conclusion

Rational Sentimentalism as a theory about the source of evaluative concepts rests on three important claims. The first claim, about the regulative role that emotions play within morality, functions to explain how a critical vocabulary of emotional appropriateness can develop for evaluative concepts composed of special kinds of emotional responses. The second claim, the commitment to basic emotions, was meant
to secure a wider sense of objectivity for the evaluative concepts under discussion. Reasons to feel gain a sense of weighty importance when they can be said to be a product of evolution necessary for human survival. The human values constituted by these emotions can then function as a check against other rational processes and provide a sense of objective guidance when instrumental reasons conflict. Although the commitment to basic emotions is now in doubt and, I bet, likely to be false, this need not spell doom for Sentimentalists wishing to make use of a more cautiously grounded empirically theory of emotion. The idea that a set of basic emotions is nearly universally shared by all normal humans was once, itself, nearly universally accepted. This has begun to change. It would be in the interest of any aspiring Sentimentalist theory of value to avoid making its soundness rest strongly on such strongly disputed empirical claims. If the roles that basic emotions are being asked to fill can be carried out without an appeal to basic emotions then so much the better for Sentimentalism. The final claim that Rational Sentimentalists commit themselves to is what I have been calling the constitutive claim. This claim will occupy my attention for the next two chapters.

The regulative role argument, I argue, marks a significant advance for Sentimentalist theories of value. The degree to which our evaluative discourse depends fundamentally upon our emotional discourse appears strong. The regulative role argument, by showing that reasons that speak to the fittingness of our emotional

---

responses have a unique kind of authority, helps to fill out an argument in favor of the emotions functioning within an epistemological role within our moral psychology. Dropping a commitment to basic emotions would not fundamentally alter the regulative role argument. It remains to be seen, however, whether it is possible to salvage Rational Sentimentalism by accounting for both the sense of normativity and objectivity that basic emotions were meant to provide. Furthermore a larger problem looms on the horizon for Sentimentalist theories of value. Sentimentalist theories of value need to argue that emotions play an irreducibly constitutive role. In the following two chapters I concern myself chiefly with canvassing for arguments in favor of the constitutive role. I begin in the fifth chapter by examining arguments offered by contemporary Sentimentalists before turning my attention in the sixth chapter to the more specific claims offered by reactive attitude theorists of moral responsibility. My claim is that these arguments are wanting. In most cases the constitutive role is either presumed or only weakly argued for and that is a problem if Sentimentalists want to offer motivation for one of their key beliefs.
In this chapter I argue that Sentimentalists have a problem. So long as Sentimentalism is understood as a theory about the ultimate meaning and ground of morality (that is, as a meta-ethical theory) then adherents need to explain why we should believe that morality is best understood as being irreducibly grounded on emotion. This argument needs motivation. I reconstruct arguments I take to be central to the constitutive commitments of many contemporary Sentimentalists and argue that these arguments do not successfully motivate the constitutive claim. In the latter sections of the chapter I attempt to learn from these attempts and, drawing from the theory of emotion developed in chapter three, attempt to offer an alternative argument. My argument takes seriously the possibility that most of our concepts operate as the Psychological Constructivist claims our emotional concepts operate. I think even this updated argument faces difficulties and, in the following chapter, turn to the reactive attitude theories of moral responsibility to support, if only in a more limited arena, the constitutive claim.

Recall the roles that emotions have been taken to play within our moral psychology. Emotions, I have said, can be asked to play three different roles within our moral framework: motivational, epistemological, or constitutive. Any moral system that accords emotion one or more of these three roles can perhaps be said to be, as Shaun Nichols has claimed, “Sentimentalist in an important sense” because they

---

1 I spoke about these roles in some detail in chapter three.

133
require some accounting of the role emotions play in a complete moral theory.\textsuperscript{2} The devil is in the details of course. These three roles posit fundamentally different relations between morality and the emotions. The motivational role, Nichols' preferred way of situating emotion within a network of moral rules, requires that emotions serve to get us to desire to act in accordance with our moral judgments. Making a moral judgment is typically, on this view, attended with moral emotions whose job it is to get us to behave morally because moral emotions are intrinsically, but defeasibly, motivating. The epistemological role goes a step further by making moral judgments dependent on possessing the right set of emotions. Emotions help us to identify, track, and understand moral situations when they are functioning in the epistemological role. The fact that I feel guilty about not spending the weekend working may in fact be an indicator of my failing to meet a promised deadline. The guilt is helping to make me aware of my moral obligations. The constitutive role goes farther still. According to first-order views, had I not felt, or thought that it was appropriate to feel guilty at missing the deadline then missing the deadline simply would not be a moral obligation if guilt is playing a constitutive role. On the constitutive role what it means to be a moral obligation depends on whether we are giving a first- or second-order analysis. On the first-order view what it means for something to be a moral obligation is that that obligation is grounded on moral emotions. If I would feel guilty at failing in my obligation then it is a moral one, if I would not then it is not. On the second-order

analysis an occurrent feeling of guilt is not required. What it means for something to be a moral obligation is for it to be true that we would judge it to be appropriate for us to feel guilty at failing to meet the obligation.\(^3\)

Since it is not quite clear what it would mean for emotions to play an 'important' role within a system of ethics, it helps to be clear what kind of demands these roles place on the relationship between emotions and ethics and, in so doing, get clearer on which role is most threatened by the codifiability problems. Some, for example, have argued that it is important to document and explain what seems to be the obvious connection between emotions and moral judgment. Shaun Nichols has recently claimed that it is enough, on his view, to argue that “we have some norms that happen to be backed by affect for a large portion of the population” because “our moral norms prohibit harming others and we have an affective system that is built to respond to harm in others.”\(^4\) This is, I argue, too imprecise to get at the specific problem I want to level against first- and second-order Sentimentalisms and indeed this is too week a role for emotions within meta-ethics more generally if they are to be properly seen as Sentimentalist. Nichols' statement is too imprecise because it is not clear how to understand what it means for emotion to 'back' a norm. Nichols seems to think that showing that emotions play the motivational role is enough to show that

---

\(^3\) Who counts as “we” is a complex question. Hume argues that any normal human that puts herself into a general point of view counts. Firth's (1952) view bears some similarities to Hume's original position. Prinz (2007) argues that everyone capable of emoting counts though their moral judgments may not agree with one another. The Rational Sentimentalists (2005) argue that all normal human beings would count because of the presence of basic emotions. Michael Smith's (1994) view also assumes all normal humans would count insofar as their idealized selves would all have similar second-order desires. John McDowell's (1985) view also counts all humans as falling under 'we.'

\(^4\) Nichols 2004 (116-117)
morality is Sentimentalist in 'an important sense.' In chapter three I argued that this was not the best way to understand moral Sentimentalism. Although my main aim in this chapter is to critique Sentimentalist theories on the grounds that they have not adequately motivated the constitutive claim, I think that being clearer about the commitments imposed by the motivational, epistemological, and constitutive roles should make clearer the essence of the Sentimentalist theory more generally. It may turn out however that these commitments, at the moment, have deeply worrying implications.

Sentimentalists, indeed all response-dependence theorists, face a challenge I have already called the problem of emotional uncodifiability. The uncodifiability problem is essentially a demand that the Sentimentalist provide an explanation or argument that supports the claim that the content of the evaluative concepts they claim are dependent upon emotions cannot instead be given a response-independent formulation. If I am right about the Sentimentalist's commitment to the constitutive thesis then Sentimentalists of all stripes must argue that moral principles derive at least some of their content from an analysis of our actual emotional responses (in the case of first-order Sentimentalisms) or the conditions that render our emotions appropriate or warranted or rational (in the case of second-order Sentimentalisms). The uncodifiability problem challenges this claim. Why not think that those things that are fearsome, shameful, or that make us morally responsible, cannot be fully explained without mentioning the emotions in an essential way? Why think that the simple
having of an emotion or that the judgment that it would be appropriate to feel an emotion is what makes it the case that something is moral or a value? Why think that emotions are playing the constitutive role? The critic of the constitutive claim has a ready analysis for many of the properties the Sentimentalist claims can only be properly understood via incorporation of emotional content: they need only point to other considerations that seem to independently justify the allegedly Sentimentalist concept.

One challenge then is to require that the Sentimentalist show that this need not be true. We should understand the uncodifiability problem as requiring more than mere coherence.\(^5\) We can demand that Sentimentalists argue not only that response-dependency is coherent but that such an analysis is necessary to properly understand some group of evaluative/moral concepts. For example we can ask for motivation why we should believe that our concept of moral responsibility should plausibly be understood in terms of feelings of resentment, indignation, or guilt. That is, we need an argument for the following:

1. Some evaluative properties are tied to emotions

\(^5\) There is the potential here for a more specific restatement of the problem that only some iterations of Sentimentalism would take on. Sentimentalists may, depending on how they understand what an emotion is, go on to add the further claim that morality itself, or what morality demands from us, is itself also uncodifiable. This would tie Sentimentalism with moral particularism in such a way as to inherit the commitments to moral uncodifiability essential to the particularist view. Although most versions of Sentimentalism (first- and second-order) do not go on to make this further particularist claim, some -the Rational Sentimentalists in particular- do seem to commit themselves to this further claim about the emotions. In the case of the Rational Sentimentalists this stems, as we'll see below, from a commitment to a non-cognitive theory of emotion. If emotions are not cognitive and uncodifiable and evaluative properties are constituted by emotion then a version of particularism results. Because this is not a universal threat to Sentimentalism I will avoid the further complication that results from particularist Sentimentalisms of this sort and focus on the arguments for the constitutive thesis alone.
2. The tie between evaluative properties and emotions is constitutive
3. The constitutive relation is irreducible

The motivational role for the emotions requires only that moral agents in the grip of an emotion are more (or less) likely to behave in certain ways. The guilty more apt to repent, the outraged more apt to fight injustice, and so on. The motivational role is often taken to imply (or be the result of) ethical *internalism*. The motivational role then, while important and independently interesting, should not, as I argued in the third chapter, be the focus of Sentimentalist's arguments when it comes to response-dependency and emotional uncodifiability. We should reject Nichols' calls to include moral theories that assign the motivational role to emotions as Sentimentalist in an 'important sense.'

The epistemological role for the emotions, recall, demands that emotions give us access to certain moral facts. In an earlier chapter I mentioned the phenomena known as 'depressive realism' which suggests that depressed or sad agents are more likely to make accurate judgments of their capabilities or prospects than happier agents. This would be an example of emotions playing the epistemological role. This role itself could have weaker and stronger formulations. Emotions could be one of many routes to moral knowledge or they may, as a result of idiosyncrasies of human nature, be the only route to moral knowledge available to us. The epistemological role, however interesting, I argued was also compatible with decidedly non-Sentimentalist conceptions of moral value. It is compatible with the epistemological role that moral facts are conceptually independent from the emotions themselves. Some moral
realists, for example, could accept the claim that emotions can often (sometimes/almost always/etc) give us access to or help us to recognize evaluative properties while still arguing that the ultimate basis for something's being an evaluative property is not the emotional response itself but instead some set of independent moral or evaluative or natural facts. While the truth of the epistemological role would add an interesting Sentimentalist flavor or intuition to a meta-ethics, it would not, for reasons to follow, satisfy the meta-ethical Sentimentalist. The epistemological role would secure the first premise but not the second or third.

The constitutive role, the role which makes evaluative concepts response-dependent, can make the emotions a metaphysically essential component of moral concepts but this is exactly what the uncodifiability objection questions. Jesse Prinz has framed the relation in the following way: “Something A belongs to the essence of another thing B if one cannot specify what it is to be B without mentioning A.” This construction is ambiguous however between A serving a strong epistemological role in terms of having knowledge about B and a truly constitutive role for A in the content of B. Just to be clear: the constitutive thesis is a claim about the content of evaluative concepts or properties (that they are irreducibly affective); the uncodifiability problem, as I am framing it here, is a question about this thesis.

Some concepts appear more amenable to a Sentimentalist analysis than others. On its face the concept of something's being funny, for example, seems difficult to explain without mentioning the kinds of things that actually (or appropriately) make us

---

6 See Prinz (2007) 18
laugh. Other concepts seem, on their face, less amenable to this kind of analysis. The concept of being *fearsome* for example appears to be intimately linked with its namesake emotion. However it does appear at least prima facie plausible to suppose that a response-independent analysis of the fearsome can be given in terms of dangerousness.\(^7\) Whether something is dangerous seems fairly independent from whether we respond to it with fear or bravado or whether we respond to it at all.\(^8\) If dangerousness does give us an analysis of fearsomeness then fearsomeness is identifiable independently of our fear responses. The job for the Sentimentalist is to explain why we should find something's being a value or a moral property to be more like something's being funny than something's being fearsome.

In previous chapters I have explained and expanded upon a theory of emotion dependent on prototype theories of concepts known in the psychological literature as 'Psychological Constructivism.' I argued that Psychological Constructivism could go a long way toward shoring up empirical issues with a particular kind of Sentimentalism: Rational Sentimentalism. Psychological Constructivism can plausibly explain emotions as we actually experience them without resorting to an appeal to basic emotions and I claimed that this theory of emotion was robust enough to account for

---

\(^7\) The Rational Sentimentalists, as I argued in the third chapter, deny that fearsomeness can be given a response-independent analysis. They claim that some things that are dangerous (I mentioned cigarettes and risky sexual practices) do not appear to be appropriate objects of fear while other things that are not dangerous (thrill rides, horror movies) invoke fear and, arguably, do this appropriately.

\(^8\) There is a trivial sense in which concepts like the dangerous are of course context relative. Dangerous *for whom*. But this it to mistake de dicto dangerous for de re dangerous. I need to be on the cliff for the cliff to be dangerous *for me* but cliffs are dangerous in the de dicto sense for all biological beings like us. This form of context relativism is very different from the constitutive role's statement that appropriate emotional responses constitute something's being a value.
evaluative disagreement of fairly broad sorts. In short, I argued for a theory of emotion that I believe may be helpful in addressing problems that many have thought made Sentimentalist meta-ethics unpalatable. What I have not yet done, however, is to argue in support of the Sentimentalist thesis more generally. In exploring the degree to which a satisfactory Sentimentalism can emerge or be constructed from the theory of emotion offered throughout the dissertation, I find myself hard pressed finding help within the literature for a positive argument for the uncodifiability thesis.

In what follows I reconstruct the strongest arguments that are available in favor of the constitutive thesis. I will also conclude that they do not go nearly far enough in terms of addressing the response-dependency question though their shortcomings help to show us one potential way of addressing the question. Since the commitment to uncodifiability is, I have suggested, essential to the meta-ethical Sentimentalist we may have expected them to respond more directly in its favor. However, aside from a few restatements or thought experiments, this is often not the case. This may seem especially problematic, it could be said, because most philosophers who posit an emotional uncodifiability thesis then go on to apparently violate it by specifying conditions, in an observer-independent fashion, of when emotions are appropriate or inappropriate, warranted or unwarranted, etc. Jay Wallace, for example, spends considerable effort to argue that our concept of moral responsibility is grounded on appropriate feelings of guilt, resentment, and indignation. However, he also somewhat

---

paradoxically argues that our concept of responsibility is tied to considerations of fairness and reasonableness. Wallace argues, for example that “moral norms of fairness have a privileged position in determining what it is to be a responsible agent.”

This kind of paradoxical response is not unique to Wallace but in fact seems to be a feature of Sentimentalist arguments more generally. It is thought that at least part of the problem lies in the commitment to uncodifiability. After all, if we reject the uncodifiability thesis then statements about the moral norms of fairness being relevant to our concept of moral responsibility, in Wallace's case, need not undermine the fact that emotions are also playing a motivational or epistemological role. In light of this issue I want to take a second look at key junctures where contemporary Sentimentalists have flirted with uncodifiability in order to see what, if any, arguments are offered.

My conclusions here will be mostly skeptical. Where arguments are necessary I find mostly promissory notes or intuition pumps. In response to this conclusion, the latter part of the chapter will take up the question of whether the uncodifiability thesis can be defended and what form a commitment to the thesis would need to take in order to save the Sentimentalist from what appears to be a less than promising argumentative commitment. Building upon the lessons gleaned from my earlier examination I suggest one potential way for a Sentimentalist theory of value built to be built up from the kind of prototype theory of emotion concepts offered in the second chapter that may be able to respond to the codifiability problem. This theory
has the potential for meeting the challenge involved with defending the uncodifiability thesis though it requires making some quite general claims about the nature of concepts and properties that may make it unpalatable to some.

5.1 Response-Dependency

What would it mean for a property to be response-dependent? I have argued that Sentimentalism is best understood as a claim about the response-dependence of moral properties on emotion. This dependence was because of a constitutive relationship that Sentimentalists posit must exist between value and emotion. Response-dependency implied more than the claim that values motivate us (to have reasons or to act) and it was not merely that we could not understand values without emotion. For a property to be response-dependent it must be an ineliminable part of the meaning of the property or concept that it is composed of or must make reference to our subjective responses. I have also analyzed two different senses, first- and second-order, in which a moral property could be thought to be dependent upon emotion.

An example of a non-evaluative response-dependent property, one with long ties to evaluative properties, is the property of being colored. Color properties, we might say, are not directly properties of objects but instead only arise in a special kind of context: one in which beings with very specific perceptual mechanisms have evolved. A property like redness, then, might be understood in part as a response-dependent property. Something is not red unless it is such as to be disposed to cause
beings like us to have red experiences when confronted with it. It is because, on this analysis, an object is disposed to cause a certain kind of experience in us that it can be properly said to instantiate a color property. Our responses, the experiences of seeing red, then are ineliminably part of the analysis of the property of being red. The claim then, what I called in chapter four the constitutive claim, is that evaluative properties function in this way. To say whether something is a value will depend on the relationship between the thing and our subjective responses. I will have more to say about the analogy with colors and the role it has historically played in motivating response-dependency about value later in the chapter.

Response-dependency has historically been seen as attractive for many reasons. Conceiving of values as response-dependent, for example, is often thought to explain some otherwise strange facts about moral judgments: that they are motivating. Non-evaluative facts, that I am in San Diego, do not carry with them any motivation (defeasible or intrinsic) to do anything in particular. Moral facts on the other hand, that it is wrong to cause wanton cruelty, seem to. In this case: do not cause

---

11 This is a superficial dispositional analysis of color properties. A more sophisticated analysis would spell out more clearly what properties of an object, the lighting conditions, our perceptual mechanisms and our situatedness it is in virtue of that something properly instantiates a color property.

12 Most response-dependent analyses of values refer to desires and not emotions as the subjective human response that constitutes values. In some cases, for example David Lewis' view, the desires are the second-order desires (desires to desire) we would have under conditions of maximum imaginative acquaintance; in other cases, for example Michael Smith's view, the desires are those that a fully informed and fully rational version of ourselves would desire our actual self to desire. For Lewis' view see his “Dispositional Theories of Value.” Proceedings of the Aristotelian Society, Supplementary Volume 63, 1989: pp. 113-137; Smith's entry in the same volume is informative as is his (1994) The Moral Problem especially his chapter “Internal Reasons.”

wanton cruelty. If moral facts were response-dependent this feature of moral facts
could be explained. Affect, as I argued in chapter three, carries with it motivational
force. Fear carries with it the command: fight or flee!\textsuperscript{14}

But not all properties are thought to be constituted by our subjective responses
even in cases where those responses can seem common or appropriate.
Dangerousness, I said earlier, is analyzable as a property that, while contextual, is
independent of subjective human responses like fear. What it means for something to
be dangerous is for it to have a certain dispositional capacity to cause humans serious
harm. A tornado is dangerous as are cigarettes as is underwater cave diving. These
things remain dangerous despite the fact that we fear some things (the tornado), do
not fear others (the cigarette) and actively seek out still others (cave diving,
cigarettes). Dangerousness appears independent of our subjective responses.
Dangerousness however can still be contextual. Technological advances may
dramatically reduce the kinds of injuries that can harm human lives. Things that are
dangerous now may not be dangerous in a future where humans are biologically
immortal. This does not change the more important fact of the matter about
dangerousness: that deciding whether or not something is dangerous need not ever
appeal to our subjective responses to objects or states of affairs. The question then
becomes whether moral properties should be more properly thought of as being like
colors or more like fearsomeness (understood as dangerousness).

\textsuperscript{14} This motivation for accepting response-dependency is, however, only as strong as the arguments for
accepting motivational internalism in the first place. For arguments against see David Brink's \textit{Moral
Realism and the Foundations of Ethics}. Cambridge: Cambridge University Press 1989. pp. 42 and
5.2 Uncodifiability: Contemporary Arguments

The uncodifiability thesis, while an important commitment for many meta-ethical Sentimentalists, is rarely given an explicit argument in its favor. Justin D'arms and Daniel Jacobson, I have argued, are committed to the emotions being uncodifiable. If the emotions are uncodifiable and moral principles get at least some of their content from the conditions that render those emotions appropriate then those principles will also be rendered uncodifiable. This is a strong version of uncodifiability that goes beyond the needs of the constitutive thesis. The constitutive thesis only requires that evaluative properties get at least some of their content from emotions. The Rational Sentimentalist agrees but adds the further claim that the conditions that render an emotion appropriate are uncodifiable in principle. Though they do not explicitly set about fully defending these commitments they say enough about them to allow for a reconstruction of an argument on their behalf. They both argue that emotional uncodifiability is a consequence of an earlier commitment to emotional non-cognitivism. The claim that the conditions that render an emotion appropriate are uncodifiable would ultimately rest on the plausibility of the arguments both for the particular brand of non-cognitivism the Rational Sentimentalists advocate along with the intuition that the inference from non-cognitivism to uncodifiability is valid. Here is how I see their argument working:

1. Emotional states are non-cognitive: there are not any constitutive thoughts, beliefs, or judgments that are essential to any emotional states.
2. Any attempt to explain or delineate the locus of concern for an emotion requires taking a non-cognitive emotional state and expressing it in terms of some set of beliefs, judgments or thoughts (about it).

3. If emotions are non-cognitive then any attempt to explain the nature of an emotional experience is 'essentially contestable'.

4. Essential contestability is a mark of uncodifiability.

5. Our evaluative concepts are essentially contestable.

6. Therefore, our evaluative concepts are uncodifiable.\(^\text{15}\)

While I have significant reasons to question the first premise, the way the third premise links emotional non-cognitivism with essential contestability is problematic.\(^\text{16}\)

Take hunger as an example. Hunger is, perhaps paradigmatically, a non-cognitive valenced affective and motivational state. Hunger is unpleasant to experience. To feel hungry is to be pained in a very particular way. Hunger is also plausibly non-cognitive. We may feel hungry without forming the thought or making the judgment that we are hungry or that we would like to eat. Hunger is also grounded in physiology and plausibly understood as a universal evolutionarily derived impulse. It makes sense to consider the adaptive advantages that a state like hunger confers upon us. A being who feels hunger is likely to survive and reproduce in greater numbers than a

---

\(^{15}\) Since D'Arms and Jacobson do not give a single canonical argument for uncodifiability per-se this extraction comes in part from their (2005) and (2006).

\(^{16}\) The problem with the first premise is simply that many emotions appear obviously cognitive. The Rational Sentimentalists argue that these obviously cognitive emotions are not the kind under analysis because they are not basic, they are instead 'cognitive sharpenings' of basic emotions. I find this to be an unsatisfactory response in the light of the arguments against basic emotions presented in previous chapters. On my view the same emotion can be both cognitive or non-cognitive depending on the particular way in which it is experienced. What matters here, what secures the 'same emotion' judgments, are the similarities an affective experience bears to the emotional prototypes. The distinction between 'basic' emotion and 'cognitive' sharpening therefore appears motivated more by the Rational Sentimentalist's non-cognitivism than by the data.
compatriot without this impulse. Hunger is also plausibly found in non-human
animals. These are all conditions that the Rational Sentimentalist thinks emotions must
have. Hunger, like amusement, seems like the kind of affective state the Rational
Sentimentalists want to give an analysis of. However, it seems that we can give a
nearly univocal gloss on the locus of concern for hunger: a need to ingest nutrients.\textsuperscript{17} If
a nearly univocal understanding of a non-cognitive state like hunger is possible then
much more needs to be said about the third premise to make emotions, as non-
cognitive states, different from other non-cognitive affective states like hunger such
that it would be plausible to believe that they would be essentially contestable in a way
that these other states do not appear to be. This is, as far as I can tell, the closest to a
direct argument for uncodifiability that the Rational Sentimentalist has given.

Jesse Prinz offers arguments in favor of a view he calls 'strong emotionism' that
mirrors the constitutive claim. Prinz defines strong emotionism as a commitment to
the following two theses\textsuperscript{18}:

\begin{itemize}
  \item \textbf{Metaphysical Thesis}: An action has the \textit{property} of being morally right
    \textit{(wrong)} Just in case it causes feelings of approbation (disapprobation)
    in normal observers under certain conditions.
  
  \item \textbf{Epistemic Thesis}: The disposition to feel emotions of approbation or
disapprobation are a possession condition on the normal concept
\textit{RIGHT (WRONG)}
\end{itemize}

The metaphysical thesis is, by all lights, a restatement of the constitutive role for

\textsuperscript{17} We may want to differentiate between various 'types' of hunger (hunger out of starvation from
hunger out of boredom etc) but even in those cases the locus of concern for hunger appears
uncontroversially univocal. If hunger doesn't convince then there are other plausible states in the
neighborhood (sleepiness, lust, thirst etc) that are commonly identified as 'drives' by psychologists
which would work in this kind of argument.

\textsuperscript{18} Prinz (2007) 20-22.
emotions under investigation here. But what arguments does Prinz rally in its support? Prinz argues that emotions are both necessary and sufficient to make it the case that our judgments become moral judgments. Moral judgments just *are* affectively made judgments. Emotions are necessary because those without access to moral emotions seem especially bad at making moral judgments, explaining moral terms, and understanding the difference between a moral wrong and a failure to follow social convention.\(^{19}\) Emotions are sufficient, according to Prinz, because very young children, those who have yet to develop significant verbal abilities, can still make judgments about the relative severity of moral transgressions with the same degree of accuracy as adults. Since children at these ages lack access to more sophisticated cognitive capacities, the likeliest conclusion, according to Prinz, is that their emotions are responsible for the judgments and hence most likely responsible for adult moral judgments as well.\(^{20}\)

\(^{19}\) Prinz here relies on data on psychopaths (42). I'll have more to say on psychopathy and constitutiveness, especially in relation to reactive attitude theories of moral responsibility, in the next chapter though see the proceeding footnote.

\(^{20}\) Some argue that the moral/conventional distinction does not map neatly onto the philosophical conceptualizations of 'moral' typically under discussion. Virtue theoretic morality, for example, does not define wrong acts as those that are independent of authority or agreement, as is the case with all studies on the moral/conventional distinction, but instead makes wrong only those actions that flow from a vicious deliberative and emotional process. I'm convinced that this is right though it casts into wide doubt our understanding of what it means to fail to make the moral/conventional distinction. See Manual Vargas and Shaun Nichols “Psychopaths and Moral Knowledge” in *Philosophy, Psychiatry, and Psychology* 14, 2, (2007), pp. 157-162. Vargas and Nichols make the further claim that a psychopath's diminished ability to identify moral versus conventional distinctions need not imply that psychopathic behavior is always excused. I think Vargas and Nichols are on the right track though will have more to say about psychopathy in the following chapter. When it comes to the case of young children that Prinz here wants to use to argue in support of strong emotionism I think this worry is doubly important. That children are able to pick up on the fact that some norms never admit of exceptions while others do may imply the beginnings of moral education, the acquisition of moral concepts, but it need not imply, as I go on to argue, that emotions are playing the constitutive role.
If this was as far as Prinz goes then it would be enough to show that the data he presents in favor of the constitutive thesis would be equally well accounted for by assigning the emotions a purely epistemological role. Psychopaths, lacking access to the moral emotions, would not be able to make moral judgments non-parasitically. Young children, having access to moral emotions, would be at least able to identify morally relevant factors via their emotions even if they could not articulate that this is what they were doing. In both cases it remains possible for moral facts to be entirely independent of the emotions being appealed to. Prinz goes on to make one further cryptic argument however.

Prinz frames his view as a response to the Euthyphro question: “Is an action wrong because we disapprove of it, or do we disapprove of it because it’s wrong?”21 The intuition here is that the second option seems to us, as it did to Euthyphro, the more attractive answer. The burden then is on the sensibility theorist to explain why it should not end up seeming bizarre that Socrates’ answer is wrong.

Sensibility theorists are committed, according to Prinz, to their being a constitutive relation between moral properties and emotions. This analysis is, as should be obvious, circular and is exactly what falls out of a commitment to the constitutive thesis. It attempts, in other words, to apparently accept both the claim that actions are wrong because we disapprove of them and that we disapprove of them because they are wrong.

But, Prinz wants to tell us, not all circularity is created equally. As a

---

21 Prinz (2007) 118.
comparison to his own view Prinz outlines an essential feature of the competing sensibility theories offered by McDowell and Wiggins. On their account they commit themselves not only to the constitutive thesis but also to a no-priority thesis. The no-priority thesis states that “there is no way to individuate the emotions that constitute our moral sentiments without reference to what those emotions designate.” Prinz thinks this is a mistake because it would make it impossible to distinguish different kinds of secondary qualities (different sentiments) from one another.

Although Prinz is right to argue that McDowell and Wiggins are not precise on what emotions/sentiments are, it is intelligible, and more charitable, to read them as holding that each emotion carves out its own sphere of concern and that our analyses of moral properties should be dependent upon an analysis of each emotion evoked by that property. Instead of adopting broad dispositional sentiments, emotion specific analyses of moral concepts (were McDowell and Wiggins to offer them) would not fall prey to Prinz’s particular objection. However, the no-priority thesis carries other costs.

Prinz intends to sidestep the no-priority thesis altogether however by offering a “way of referring to moral emotions without referring to that [about] which we moralize.” How does he do this?

He imagines that certain actions “not yet specific to the moral domain” elicit emotional responses from us. We may feel angry, for example, when someone does something cruel to us and this anger may cause us to engage in anti-social behavior

\[\text{Ibid.}\]
\[\text{Prinz (2007) 118}\]
toward the perpetrator. Once patterns of behavior like this “become stable” then moral rules have come to govern over a certain kind of behavior in virtue of being able to elicit emotions from us. As Prinz says, “guilt and righteous anger are born.”

Prinz thinks that this response, though still circular, avoids the problems that come with a no-priority thesis because we can specify why we find some particular action to be morally wrong. We can say, for example, that something is wrong because it is immoral and it is immoral because it violates a person’s autonomy and we find this infuriating. At bottom the moral rule is based on our emotional responses but the additional content given to the analysis via the route through moral rules is supposed to give the account an explanatory power lacking from the no-priority views.

I admit to being a bit confused about why we should understand the theory here as helping convince us of the truth of the constitutive relation between evaluative properties and, in Prinz's case, sentiments. The genealogy presented appears to presume the truth of the constitutive claim more than it argues for it. Since Prinz has taken himself to have shown that emotions play a necessary and sufficient role in morality and that this role is understandable in terms of a constructivist process then it stands to reason, one might say on Prinz's behalf, that the fact that changes in making moral judgments and society-wide level changes in moral values are accompanied by changes in our emotional lives that these facts are best explained by the constitutive relation.

---

24 Ibid. 118-119
25 David Wiggins offers a theory of evaluative concepts, he calls it “Sensible Subjectivism,” that can help fill out one version of the codifiability problem. Wiggins, like Hume, attempts to give a vaguely etymological account of evaluative concepts where we begin with simple reactive
Prinz offers one more argument for the response-dependency of evaluative concepts. His argument is focused on a particular kind of analysis of our folk evaluative concepts and comparing those concepts to other, similar, concepts involving our folk understandings of color properties. Below is a reconstruction of that argument:

1. Folk color concepts define color properties as mind-independent properties of objects

2. Color properties are not mind-independent properties of objects, they are response-dependent properties

3. Therefore folk color concepts contain errors about the nature of color

4. If a causal theory of reference is true then concepts can contain errors while successfully referring to the properties that actually cause color experiences in us

5. A causal theory of reference is true

6. Therefore color concepts, despite their errors, still successfully refer to the response-dependent properties that actually cause color experiences in us

Without an argument like this Prinz says that our only recourse, in the face of errors in our concepts, is nihilism about the erroneous concepts. We would have to sacrifice our color concepts. Because of the very real roles that color concepts play in our lives (aesthetic, descriptive, pragmatic, etc) it makes more sense to keep the concept but abandon some of the errors built into them at the folk level. For Prinz the error of folk color concepts is that color properties are characterized as being mind-independent.

---

This is, according to Prinz, false. Colors are not properties of objects or surfaces, they are relational. Because Prinz relies on Fred Dretske's causal theory of reference, our concepts refer to the things that reliably cause us to invoke the concept in question; color concepts can still refer to the response-dependent properties of surfaces and brains that are actually responsible for our color experience.\(^{26}\) Although this means that we will have to change how we understand what colors are, they are not mind-independent properties of objects, we can save the discourse and the practices the discourse supports and have that discourse better reflect the true nature of our concepts.

The argument, then, is that evaluative properties can be explained analogously. Our folk evaluative concepts contain errors. In particular our folk evaluative concepts define evaluative properties as mind-independent properties of the world. Given that Prinz is a first-order Sentimentalist he believes that this is a mistaken assumption about the nature of moral properties. Furthermore Prinz believes that cross-cultural data on moral judgments imply that moral properties vary with culture.\(^{27}\) Indeed because his is a first-order view they may vary from person to person. But moral discourse, much like discourse about color, plays many different and important roles for us. Moral nihilism then would be a bad option to take in the face of these errors in our evaluative properties. Moral concepts are reliably invoked, for Prinz, by our

\(^{26}\) In a similar way Prinz argues that our concepts for Gorilla, to give another example, can still successfully refer to gorillas even though they often contain many errors (that gorillas are aggressive or dangerous).

\(^{27}\) I focus more on cultural differences in the next chapter. There I suggest that important differences may exist between guilt-cultures and shame-cultures that lend credence to Prinz's argument here.
sentiments and the sentiments are sets of emotional dispositions. This means that our moral concepts, and the moral properties they invoke, still successfully refer to the response-dependent properties that cause them.

If Prinz is right about this approach then response-dependency is vindicated. But how far afield does this argument take us? If we had sought out to motivate response-dependency we should not, as a key premise in that argument, assume against mind-independence. What motivated the key premise against mind-independence? Prinz is impressed by the empirical data on the moral structures not just across different cultures but also within different groups in the same culture. Prinz argues, for example, that liberals and conservatives in the United States may be unable to reach agreements on political values because their political beliefs are based on different grounding norms. Since grounding norms form the base of all moral judgments this fact means, for Prinz, that liberals and conservatives waste their time debating with one another. The best way to account for those differences is via an appeal to his collected data on emotions and the role they play in moral judgments. This 'argument from variation' relies on the claim that different groups of persons see the same situation as differently valenced. Because it sounds odd to say that the same situation is both moral and immoral, funny and not funny, and so on, the best explanation for different groups evaluating a situation differently is due to the situation itself not being one way or another without specifying a context of evaluation. For

28 His solution here is that we may have to resort to non-evaluative or moral means in order to change the minds of those with different grounding norms (we may have to resort to “the fist” to use his terms). See Prinz (2007) 125.
example we could observe that it is a fact that its being '85 degrees Fahrenheit' is, for coastal San Diegans, quite hot but also that it is quite comfortable for dessert bound Phoenicians and that this dichotomy implies that its being 85 degrees Fahrenheit is neither hot nor cold unless we specify a judging context. Something like this seems behind Prinz's inclination to see difference in moral concepts across time and place as implying relativism.

But this assumes the very premise at issue here. The mere fact of disagreement should not lead us to conclude that there is not a fact of the matter, at least not in most contexts. If I claim that there is an elephant in the room and you do not see an elephant it seems right to say that one of us is right and that what makes us right is that we are better situated to make those judgments, for example that we are not under the influence of any medications or otherwise deluded. In the next chapter I will take a closer look at another, more successful, proposal to use cultural differences to ground response-dependency at least about the narrow class of emotions called 'reactive attitudes' and the role they play in conceptions of moral responsibility. But here I conclude that Prinz's appeal to an analogy between emotion and color and to suggest a link between arguments for response-dependent theories of color and Sentimentalism is too quick. We need some reason for thinking that different groups are equally well situated to make moral judgments or that there is not a plausible situation from which those judgments are better made.

Furthermore the argument that there is a meaningful analogy between color
and value has a troubled history. Simon Blackburn, for example, has argued that colors and values are disanalogous in several morally relevant ways.\textsuperscript{29} Our judgments about color are not, he argues, objective in the same sense in which we take our moral judgments to be. Inverting our color spectrum may, according to Blackburn, end up changing the color properties of an object (what was red is now green) but moral properties are not thought to vary in this way. Changing how we feel about murder does not change the wrongness of murder, Blackburn might say, despite Prinz's attempt to claim that it should. Blackburn argues that our moral concepts are unlike color concepts because they are rigid. Their extension does not vary with our sensibilities as color properties appear to vary. The rigidity of our moral concepts has itself been challenged. Peter Railton has argued that rigidifying color concepts makes sense given the function of color concepts in color discourse but that the rigidifying strategy is a mistake for moral concepts.\textsuperscript{30} Although Railton agrees with Blackburn that colors make for a bad analogy with values he is led to claim instead that this should push us to look for other secondary qualities that make for a more apt analogy with values.\textsuperscript{31} Railton argues that gustatory qualities are better analogues for evaluative properties. Something's being a value is more like something's being bitter


\textsuperscript{30} David Lewis has also questioned whether rigidifying our evaluative concepts really manages to respond to Blackburn's concerns over relativism. Rigidifying our actual evaluative concepts still leaves us without a principled reason to think that we could not have had some other actual set of evaluative concepts instead of our current set. It would appear a strange coincidence to assume that we necessarily must have had the set of values that we currently have and hence the status of our evaluative concepts is left threatened even if we rigidify them. See his "Dispositional Theories of Value" in the \textit{Proceedings of the Aristotelian Society, Supplementary Volume} 63, 1989: pp. 113-137.

than something's being red. In order for something to be a value it must at least in principle be recognizable as a value by an agent. If agents were to change such that they were incapable of recognizing something as a value then it would cease to be a value for them. Blackburn's inverted color spectra on the other hand has led many theorists of color properties to disagree with Blackburn's initial analysis. They do not believe that inverting our color experiences would necessarily change the color properties of objects. They argue that we should rigidify color properties to actual human sensibilities and standard lighting conditions to avoid relativizing color properties. On their terms the color analogy seems in line with Blackburn's intuitions about the rigidity of moral concepts. Railton thinks the analogous move with values is a mistake.

Although the arguments examined thus far do not succeed in vindicating the constitutive claim some progress has been made. In the final chapter I look more closely at cross-cultural moral data to extend Prinz's argument from variation. For the moment I conclude that Prinz's appeal to an analogy between emotion and color and to suggest a link between arguments for response-dependent theories of color and Sentimentalism is too quick.

5.3 Beyond Uncodifiability: Explaining Evaluative Discourse without Basic Emotions

Although Prinz's move may seem to pass too quickly, I think that he is

---

32 Railton's example here involves imagining a future race of cloned humans who lack genetic kin preferences: they do not care intrinsically about the lives of genetically related persons any more than they do about other persons they may have relationships with. In a world made up entirely of these clones Railton argues that it makes little sense to say that there is something distinctly valuable about kin interactions even though we non-cloned humans do distinctly value kin interaction.
importantly on to the right track. If Sentimentalists want to secure an argument for the constitutive thesis and hold fast to uncodifiability then I think a shift toward constructivism is a shift in the right direction. Constructivism is poised, for example, to explain essential contestability in general terms. What I mean by this is that constructivism, as a general theory of meaning, can adopt moral language as a special case of the more general project of understanding meaning. Some concepts may be more or less tied to observer-independent properties and hence more or less malleable in terms of social influence over their meaning but the critical point remains the same: intersubjective demands on our concepts can fundamentally alter them. If these demands are in part being guided or initiated by emotional responses then the emotions are functioning in what looks like the constitutive role in these specific instances. This would tie the prospects of the resulting Sentimentalism with the prospects of constructivist arguments generally and, while a risky commitment, may be worth the price of admission if it provides us with the arguments in favor of the constitutive thesis and emotional uncodifiability that are lacking. In the spirit of this enterprise I want to see how far a view like this can be taken.

David Wiggins and Jesse Prinz both offer genealogical stories meant to illustrate the origin of evaluative concepts and make the relationship between values and emotion clearer. The story is of course subject to empirical falsification but the appeal of offering a theory like this is in the fact that they satisfy a need that subjectivists generally (and Sentimentalists particularly) must meet. They need to be
able to explain why our moral concepts are constituted by emotions, but this can take more than one meaning. Part of the story lies in giving a conceptual account of what it would mean for a moral concept or an evaluative property to be response-dependent (this is where much of the argument takes place). For D'Arms and Jacobson this means explaining what it is about appropriate emotions that makes them unique. For the Rational Sentimentalists the answer lies in the fact that appropriate emotions respond to a particular emotion's locus of concern. This is something that neither pragmatic nor etiquette based concerns consistently do. They must also explain why this fact about emotions is important. A consideration about the the appropriateness of an emotion serves an important regulative role in our emotional lives and this in turn helps to explain much of our emotional discourse. It explains why we try to calm ourselves, if we fear flying, by focusing on the safety of air travel or why we might convince ourselves that we have not actually done anything wrong in order to soothe our guilt. On the other hand, genealogical arguments matter for Jesse Prinz, David Wiggins and Jay Wallace because they want to be able to explain cultural differences in evaluative practices, concepts and discourses and they want to explain them as resulting from differences in each culture's emotional repertoire.

David Wiggins offers a genealogical account of the origin of anthropocentric evaluative concepts that opens by imagining that we first begin with unrefined pairings of subjective responses and different features of the world and ends with an account of the evaluative discourse these pairings eventually develop into. In what

33 Wiggins (1987)
follows I want to explain the contours of this approach in order to show how subjectivists have appealed to genealogical stories of concept formation to secure response-dependency. These stories, if true, carry with them potential costs. They may generalize too strongly by making all concepts broadly response-dependent.

We are not here positing innate moral knowledge. As a result we must somehow explain how we form evaluative/moral concepts from particular experiences. If every individual relied solely on her own experiences to develop moral concepts there would be an immediate set of problems. Her moral concepts would be idiosyncratic (because they are solely her own) and unstable (because her experiences would be in flux). I earlier referred to this as Hume's Problem. Two things are important here:

1. What we are seeking to do now is to make stable judgments about our moral impressions/moral judgments.

2. Achieving (1) requires that we somehow acquire for ourselves a more ‘steady’ and ‘general’ revival set for our moral concepts

Both goals are achieved via moral discourse. As we share the particulars with another to which we have reactively responded with what we might call bare-concepts we stabilize these concepts of evaluative properties by shifting the standard of correctness for a concept's application from a personal level to a public level. The realization that others have different extensions for moral terms and that our own moral concepts are in a state of flux due to our constantly changing particular situation is a cause for concern. We correct these impressions as a result of our search for more stable
concepts. For example David Hume remarked that time and distance affect the strength of moral impressions (a murder that occurred eons ago excites our disapprobation far less than one that occurs before our very eyes though we would admit and judge that both are equally morally wrong). Furthermore our own experience is limited and varied such that sharing accounts of our experiences with others can serve to broaden our understanding of our own moral faculties. It is necessary therefore to acquire moral concepts via public discourse in order to be able to acquire steady and general (more useful) concepts.

If Sentimentalists are to provide an account of evaluative discourse that does not rely on basic emotions then they need a story that explains how evaluative discourse is built up from emotions that are not 'basic' in the technical sense I have been describing. If emotions are viewed as constructed from core affect along with other components which make up the prototypes and if at least some evaluative discourse is based upon these concepts then a theory of the kind on offer here becomes more relevant, and useful. The take home point is that the foundations of moral concepts are what we might call bare concepts. Bare concepts, being bare, are non-propositional and (most likely) non-conscious. Public discourse is necessary for the acquisition of moral concepts and by analogy other concepts are as well.\textsuperscript{34}

On this point then the Psychological Constructivist's story about the origin of moral concepts are offered by David Wiggins (1987) and Jesse Prinz (2007). Both rely on assumptions about what makes a moral judgment meaningful as a moral judgment (the response or emotion involved) and then give a broadly constructivist account of how that meaning comes to be associated with their specific vehicles in such a way that truths within that domain are subjective.
our emotional types meshes with the prototype theorists explanation of concept formation more generally. It is not up to me, all by myself, to determine whether someone is morally responsible for her actions, for example. Whether someone is responsible seems to be a fact about the agent and the world and that fact has an objective character. The account offered above explains why this intuitive understanding of our moral judgments appears true. We really cannot, by ourselves, determine ex nihilo whether an agent is morally responsible for her actions. That requires acquiring the public, intersubjective, concepts. If these concepts are grounded on something like David Wiggins' bare pairings of emotional response and property then we have a Sentimentalist theory of these concepts. As we will see in the final chapter, moral responsibility is thought by some to be explained by a combination of emotional responses and the conditions for their appropriateness in just this way.35

This feature of the theory helps to explain the function of judgments about an emotion's appropriateness in the Rational Sentimentalist's theory of evaluative concepts. Appropriate emotions are those that closely match the prototype or exemplar for their types. In James Russell's sense, appropriate emotions are those that most closely match the 'script' for a particular emotion. Other emotions are judged by reference to how closely they follow this script.

The nature of the constructive process means that the concepts and responses involved are what some have called 'essentially contestable'; the concepts are, at the

35 More likely, I argue, is that essential components of a theory of moral responsibility, like attributibility or an account of agency, may be better places to look for Sentimentalist arguments in favor of the constitutive claim.
public level, always open to change as our understanding and uses for different kinds of concepts changes. In the ideal case the Sentimentalist would now be in a position to argue that essential contestability, now made plausible, is a major step toward responding to the codifiability concern. If it can be shown that emotionally based evaluative concepts are essentially contestable, that there are not canonical statements of the intensional content of evaluative concepts, then it is plausible to argue that this fact—assuming it is a deep fact about our evaluative concepts—helps to show that evaluative concepts can be coherently described as being uncodifiable.

I argued that there were two ways of conceiving of the uncodifiability worry. The first was a demand to provide a motivation for response-dependency. On this view an evaluative concept should be considered codifiable if it is possible to account for the content of the concept without making references to the property (or the responses it elicits). This would require us to be able to account for the content of the concepts in an observer-independent fashion. Since, on the accounts under consideration at the moment, evaluative concepts are the result of both:

1. Emotions which are themselves built up via a partially constructivist process
2. Public (moral) discourse about the concepts themselves

there should not, according to the theory on the table, be a canonical or observer-independent formulation of the content of an evaluative concept because the formulation is constituted at the social level by a fairly univocal (though always up for grabs) stable social understanding of not only what emotions there are but what those
emotions are about. This means that the evaluative concepts built up from emotions in the way I have been describing would be codifiable but dependent on subjective emotional responses. This means that an argument for the constitutive role is possible. On the other hand I also suggested that the constitutive thesis is given a different reading by the Rational Sentimentalists: a particularist reading. The particularist claim is that morality itself cannot be codified into principles or rules. Although this claim, I said, is defended by the Rational Sentimentalists as a result of their emotional non-cognitivism it need not be a feature of Sentimentalism generally. If emotions constitute public moral concepts in the way described above then they may be codifiable as moral principles, Sentimentalist moral principles. The Rational Sentimentalist therefore supports particularism on the basis of emotional non-cognitivism but the argument for emotional non-cognitivism was not convincing. It relied on the existence of discrete physiological patterns that made emotional states independently verifiable. These states, I claimed, do not exist. This version of the uncodifiability worry remains a live problem for the Rational Sentimentalist.

Even in the best case however the resulting normative standards may not be able to address all the demands that we ask it to meet. They may not be able to answer to all possible criticism in which case the standard itself will have to be replaced but this is not inevitable and it need not be a threat to subjectivism.\(^{36}\) In cases like this, Wiggins argues that

\[^{36}\text{See Wiggins 1987. p. 201}\]
aesthetic worlds whose inhabitants need to struggle long and hard to appreciate the differences. They must come to understand not only the nature and extent of these differences, but also something of the way in which these differences are shaped substantively by the conditions of human life, before it is even an option for them to come properly to disagree with one another in their valuations.37

This is the kind of result we should expect if the prototype account developed earlier is true. One of the upshots of the theory on offer here is that it fits with data that show that large differences between cultural groups (think back to the differences in emotional identifications and convergence between western groups and isolated illiterate non-western groups first explained in chapter three) are paralleled by large differences between how those groups carve up affective space into discrete emotions. This would imply that these groups would indeed have to work quite a bit in order to fully understand the emotional lives of others. If it is plausible to think that evaluative concepts are grounded on those emotional responses then it makes sense to say that they must come to understand not only the nature and extent of their affective differences, but also something of the way in which these differences are shaped substantively by the conditions of human life. They must do this before it is even an option for them to come properly to disagree with one another in their valuations.38

This may, on its face, appear to be a problem for the view.

37 Ibid, 203
38 Peter Goldie (2000) offers what he calls a 'real life' example of this phenomena. He argues that the emotions of fago and gafago felt by “the people of the Ifaluk atoll in Micronesia” seem to combine many different western emotions together. Feelings of fago are owed to individuals in the grip of gafago though the exact nature of these emotions are difficult for Westerners to understand. Goldie argues that an an anthropologist, not equipped with either emotion, may come to “gain a grasp of both concepts in a away which is sufficient reliably to understand the cultural and ethical evidence of the recognition-response tie for the islanders” though this would require a concerted effort (32).
The problematic implication being that we would not be able to criticize different persons for actions that, to us, appear immoral. This begs the question against the subjectivist premise. Part of the point of appealing to subjectivism is to explain why evaluative properties differ among groups but also to explain why moral discourse within groups can change moral understanding. The fact that an implication of this kind of view is that moral discourse between groups is constrained by the degree of similarity of their moral concepts should not only be expected, it should be accepted as the best explanation of moral discourse and moral concepts. So says the subjectivist anyway.\textsuperscript{39}

5.5 Uncodifiability

Where does this account leave us then in terms of an argument for emotional uncodifiability and the constitutive role of emotions? What kind of argument has been provided and how strong a case can be made that our moral concepts are irreducibly affective? At the outset of the project I stated my thesis interrogatively: Can an account of emotion and Sentimentalism be offered that addresses both the empirical problems inherent in the philosophy of emotion while also addressing philosophical questions regarding the uncodifiability thesis?

The meta-ethical Sentimentalist needs to defend the constitutive role for the emotions. Only the constitutive role assigns to emotions a role central enough to for emotions to support what Prinz called the 'metaphysical thesis.' While the Rational

\textsuperscript{39} I consider a different approach to other cross-cultural norms in the following chapter. There I consider appeals to cultural differences in reactive attitudes to look for corresponding differences in moral concepts.
Sentimentalists are betting that essential contestability can be secured via an appeal to a non-cognitive theory of emotion I have tried to suggest that even if non-cognitivism is true that essential contestability need not follow. While Prinz, I believe, is on the right track in terms of attempting to offer a constructivist explanation of moral properties the story itself seems to rely on the truth of the constitutive thesis as the best explanation for our moral discourse and moral practice.

I have argued that there is empirical evidence for a constructivist theory of emotion. I supported the claim that, given the prototype nature of our emotional concepts, that moral discourse requires inter-subjectivity in order to create a stable set of affective concepts (emotions). This was the case because of the need to stabilize the concepts. It is therefore a viable option for Sentimentalists to construct a theory of moral concepts from the underlying emotional framework which would in turn inherit important features of its emotional progenitors. This goes some distance toward supporting the Sentimentalist position; farther, I argue, than many contemporary Sentimentalists have gone while arguing in favor of uncodifiability. The argument does have its limitations however.

One limitation is the degree to which a constructivist genealogy can be said to simultaneous prove too much and simultaneously to not prove enough. It may be said to prove too much insofar as it suggests that all of our concepts are essentially contestable and possibly response-dependent. On the other hand some might question whether simply showing that our conception of a thing is essentially contestable
proves that the thing itself is response-dependent. There are many different competing
conceptions of the good, for example. Some of these conceptions are response-
dependent and some are not. That the Sentimentalist can deliver a conception of the
good that is response-dependent does not guarantee that goodness itself, at the end of
the day, is response-dependent in the way the Sentimentalist suggests. Fearsomeness, I
noted, can be conceived of in both response-dependent and response-independent
ways. More needs to be said to show that the concept itself takes the response-
dependent form. These are important questions to which I can only begin a cursory
discussion. These critiques threaten to render the Prinz/Wiggins approach otiose. If,
for example, one wants to argue that only evaluative concepts are essentially
contestable then a further argument will be needed that explains what makes them
different from other concepts. This would render the genealogical/constructivist theory
redundant since the real philosophical work would then take place with whatever
argument renders the evaluative concepts special.

This is a considerable further demand. On the one hand one could bite the
bullet and agree that all concepts are essentially contestable in the way the
constructive account suggests. This need not spell doom since important differences
may still exist between kinds of constructivist concepts. The constructive theory of
emotion I have offered, for example, is constrained by its tie to core affect. Core affect
limits what could count as an emotion and hence the kind of conception of an emotion
that could be constructed.\textsuperscript{40} Other seemingly mind-independent concepts, like our

\textsuperscript{40} This was one why James Russell called his theory 'Psychological Constructivism' instead of simply
concept of a star, would have to, if we are biting bullets, be constrained in part by some features as well, perhaps by limitations on human conceivable.\footnote{\textsuperscript{41} Further evidence for the view that concepts and the objects they pick out have soft boundaries could be found, perhaps, by looking at the intense debate among astrophysicists about the lower limits for stars and the upper limits for gas giants. Brown dwarfs, for example, have been classified as both gas giant planets and as low mass stars. This might look like a form of Sorites paradox on the stellar level but it is not. The debate here centers on what differentiates a true star from a non-star (gas giants, brown dwarfs, and so on). It makes clearer the fact that the concept of a star may very well be best understood as a construction, one in which there is currently active debate about whether brown dwarfs lie close enough/far enough away from the prototype to count as an instance of the type. For a review of the brown dwarf debate see Gibor Basri and Michael Brown's "Planetimals to Brown Dwarfs: What is a Planet?" \textit{Annual Review of Earth and Planetary Science} 34 2006: pp. 193–216.} Emotions are similarly constrained, as I have already suggested, by core affect. Within the space of core affect the conceptual space is up for grabs and intersubjective demands and cultural history play an important role. But ideally we would not have to bite this bullet. The ideal argument for the constitutive thesis would set a principled reason why some concepts seem bound by mind-independent facts while others are response-dependent. In the next chapter I turn to a narrower series of arguments based on the reactive attitude tradition in theories of moral responsibility in order to see whether something approaching this ideal argument can be found.

Insofar as the constitutive thesis itself should be taken to be true in light of specific features of moral discourse or other sets of standing moral commitments then I believe that, at a general level, there are reasons that speak in favor of it but, as I noted above, there are also good reasons for wanting a stronger argument here. In this chapter I offered a new way of trying to meet the demand for an argument for the constitutive thesis. The view, dependent on a psychologist constructivist analysis of
emotional concepts, was able to secure the constitutive relation. This security however came at a price. All, or nearly all, concepts would have to be essentially contestable, constructivist, in order for the argument to carry through. While I offered reasons for thinking that even some of our astrophysical concepts seem to behave in line with this constructivist analysis, Sentimentalists would, ideally, produce an argument with less obviously revisionist implications. For at least some evaluative properties (the humorous, the shameful, the fearsome) a strong plausibility argument exists. In many cases however Sentimentalists have not gone far enough in terms of arguing in favor of uncodifiability. Narrowing the subject matter from evaluative properties generally and emotions generally to the smaller classes of 'moral responsibility' and 'reactive attitudes' may lessen the argumentative burden here and point the way toward a better form of argument for the constitutive claim. I turn my attention to these issues in the following chapter.

Chapter 5, in part, is being prepared for publication in *Philosophical Psychology*. The dissertation author was the primary author of this material.
Chapter 6

Moral Responsibility, Reactive Attitudes, and the Constitutive Claim

Chapter five ended on a cautionary note. Sentimentalists, I argued, need to find a way to motivate and argue for the constitutive relationship they posit exists between emotion and value. Many Sentimentalists are driven by the intuitive appeal of constitutiveness but few arguments are available in the literature. Response-dependency, we said, was a thesis about the relationship between a concept or property and subjective human responses. Some properties, color properties or gustatory properties, seem especially amenable to an analysis in terms of our responses. Other properties, like something's being fearsome, are intimately bound up with subjective human responses like fear but are plausibly not defined in terms of those responses. A fearsome thing remains fearsome whether or not anyone is ever afraid of it or even if everyone agrees that it is appropriate not to feel fear toward it. So long as it actually has the capacity to cause serious harm, so long as it is actually dangerous, then it is plausible to call it fearsome. Response-dependent analyses of values, we said, analogize evaluative properties with color properties. There were reasons for doubting the aptness of this analogy. Advocates of response-dependency argued that other secondary qualities, gustatory ones, proved better analogues for evaluative properties because they are not rigidified like color properties. Detractors make the opposite claim. Color properties are not objective enough to capture our use of evaluative terms. In both cases it seems the analogy with color is problematic.

I turned my attention to one of the leading contemporary iterations of
Sentimentalism, Rational Sentimentalism, for an argument. I said that the Rational Sentimentalists aim to secure constitutiveness by appealing to a theory of emotion that was both non-cognitive and basic.¹ In the previous chapter I raised two worries for this kind of response to the uncodifiability worry. I argued that emotions are not basic. We do not have good evidence for thinking that emotions are basic in the sense appealed to by these Sentimentalists. Furthermore there are good reasons for thinking that emotions are not always non-cognitive.² If emotions are not basic or necessarily non-cognitive then much of the intuitive appeal of the Rational Sentimentalists ability to explain moral discourse is lost. I offered the psychological-constructivist approach as one way to begin ameliorating this worry about evaluative disagreements and non-cognitivism. Secondly, the argument from non-cognitivism to constitutiveness was incomplete. Non-cognitivism was taken to imply uncodifiability and uncodifiability was meant to imply constitutiveness. There were two problems here. Non-cognitivism, at least as a catch all theory of emotion is also false. Emotions, I argued, were best understood as prototypes admitting of emotional episodes that were cognitive and non-cognitive. Further, even if true, it was not obvious that non-cognitivism implied uncodifiability. The mere fact of non-cognitivism is compatible with emotional responses being codifiable just as other non-cognitive affective states like hunger are

¹ Recall from the introductory chapter that my use of non-cognitive here in reference to emotion is distinct from non-cognitivist theories of moral judgments. Emotional non-cognitivism only implies that an emotional episode does not necessarily require propositional attitudes (beliefs, judgments, etc) whereas the meta-ethical variety is a claim about whether moral utterances have propositional structure (whether they are truth amenable).

² I also said that Rational Sentimentalists inherit the further particularist problem that results from combining their non-cognitive theory of emotion with Sentimentalism. The problem here, if it is fair to call it a problem, stems from the resulting commitment to moral particularism.
codifiable. Other Sentimentalists appealed to early childhood moral development in the hopes of showing that affective capacities are playing the constitutive role. Here too I argued that the evidence can be explained by appealing to epistemological function for the emotions. Constitutiveness remained elusive. I closed by offering an alternative argument for the emotional constitution of evaluative properties. Following the approaches of Jesse Prinz and David Wiggins I considered the plausibility of genealogical explanations for moral concepts. These implicated our need for intersubjective communication in the development of concepts to explain the development of the kind of affectively laden evaluative concepts Sentimentalists are most concerned with. The primary function of those concepts was to demarcate important pro-social norms. These views would explain why affect and the emotional concepts that are built up from affect constitute, by contributing content to, evaluative properties. But this view comes with a heavy price: it rendered all concepts essentially contestable. Although this need not be a damning worry, an ideal response to the uncodifiability worry would give us a more principled basis for drawing a boundary between moral concepts and other concepts.

In this chapter I aim to improve upon this view. Instead of seeking a global account of the emotional constitution of evaluative properties I aim to narrow the analysis. I focus in this chapter on theories that aim to explain a much smaller set of evaluative concepts as constituted by an equally narrower set of emotions. A major family of theories on the nature of moral responsibility received renewed interest with
Peter Strawson's publication of “Freedom and Resentment.” These theories are known as reactive attitude theories of moral responsibility. Reactive attitude theorists argue that what it means for someone to be a morally responsible agent is for them to feel and be open to being targets of some set of emotions. These emotions are the reactive attitudes. On the schematic laid out in previous chapters reactive attitude theorists are Sentimentalist theories of moral responsibility. They hold that the relationship between emotion and responsibility is not merely motivational or epistemic but constitutive. The concept of responsibility is, at least in part, defined by the conditions that make the reactive attitudes appropriate and thus these are second-order Sentimentalist theories.

In this chapter I examine this family of theories to assess arguments for the constitutive role. Reactive attitude theorists go some way toward motivating the constitutive thesis but here again arguments are slim. Appeals to our practice of holding others responsible and appeals to variability in shame focused and guilt focused cultures may go some distance toward motivating constitutiveness but not far enough, I argue. Appeals to aberrant agents, psychopaths, though often thought to prove that emotions are constitutive of moral concepts serve much better as providing support for the epistemological role. Here again I conclude that reactive attitude theorists do not yet have the resources to stave off the objections to the constitutive thesis though much as I did in the last chapter I aim to improve upon these arguments.

I end the chapter by proposing an empirical test for reactive attitude theories of
moral responsibility. Although appeals to psychopathic agents by themselves do not show that emotions are playing a constitutive role I argue that they can help form an interesting burden shifting argument in favor of the constitutive claim for the reactive attitudes and moral responsibility. By combining the psychopathic data with cross-cultural data on what are sometimes called 'shame' cultures it is possible to argue that the reactive attitudes are doing more than motivating us to make responsibility ascriptions and that they are doing more than giving us access to response-independent conceptions of moral responsibility. The argument stems in part from the view that some reactive attitudes: guilt, resentment, and the like, appear to have moral content as an ineliminable part of what the Rational Sentimentalists would call their 'locus of concern.' Guilt for example seems to be a feeling that arises from the judgment that one has acted wrongly. If shame-cultures are possible then those cultures would use other sets of reactive attitudes to enforce pro-social behavior. Those reactive attitudes would not include moral content as an ineliminable part of their locus of concern and hence shame-cultures will form, if the constitutive thesis is correct, a correspondingly different set of normative concepts backed by a different set of reactive attitudes. This would be suggestive evidence for constitutiveness. Or so I will claim.

I will draw on some theoretical arguments by Jay Wallace, Bernard Williams and others along with cross-cultural data on reactive attitudes to suggest that this form of argument deserves further development. Data is scant but promising. The empirical case for constitution, at least for reactive attitudes, is possible and this way of
supporting the constitutive claim goes further than other contemporary views have in terms of providing support that motivates taking constitutiveness seriously. Some claim that only guilt-based cultures have distinctly moral concepts because of the relation between guilt and the concept of wrongness. Finding empirical evidence that shame-cultures lack the kinds of practices supported by guilt and the distinctly moral sense of wrongness invoked by guilt would go some way toward supporting the claim that reactive attitudes like resentment, indignation, and guilt constitute the family of moral concepts related to moral responsibility.

6.1 Moral Responsibility and Reactive Attitudes

Philosophers debate about moral responsibility along several dimensions. One important dimension is an analysis of agency: in virtue of what is something an agent? Furthermore, what makes those agents responsible for (at least some) of their actions? One topic that has historically been of critical importance is the question of whether our concept of responsible agency is compatible with the truth of different forms of determinism. Psychological determinism is the claim that our judgments are entirely determined by prior thoughts, beliefs, desires, or judgments. That is, an agent could

---

3 Gary Watson's influential “Two Faces of Responsibility” in *Philosophical Topics* 24 (2) has had a lasting impact on theories of moral responsibility. Watson distinguishes between two different senses of responsibility: attributability and accountability. Attributability is the sense in which an action can be said to properly belong to and reflect upon an agent. Accountability is the separate matter of addressing an attributable agent in different ways (assessing them as an appropriate target of the reactive attitudes, actually addressing them reactively, punishing, and so on). For a useful discussion of the increasingly complex taxonomy that has come in the wake of Watson's analysis see John Martin Fischer and Neil Tognazzini's “The Physiognomy of Responsibility” in *Philosophy and Phenomenological Research* 82(2) 2011: pp. 381-417. In my terms the reactive attitudes are potentially constituting both the conditions that render an agent attributable (what makes it the case that an action adheres to an agent) and whether some form of moral address or punishment is merited. Other concerns, especially consequentialist concerns, become relevant in the matter of institutional punishments however.
not do anything other than what she did given certain facts about her psychology and
certain facts about human psychology more generally. Physical determinism is the
claim that a full description of a given state of the universe coupled with a true
account of the laws of nature is enough to predict any past or future state of the
universe. Physical determinism implies psychological determinism but the reverse is
not true. The question at hand for many philosophers is whether we can be responsible
given the truth of either form of determinism. Compatibilists about moral
responsibility answer this question affirmatively. Incompatibilists respond negatively.
The disagreement between the two groups can be characterized as consisting in
conflicting analyses of the concept of responsible agency. Incompatibilists argue that
responsible agency is best understood as requiring a certain kind of power: the power
to do something other than what our prior beliefs, judgments, desires or the laws of
nature decree. Compatibilists offer competing analyses of responsible agency that
ground responsibility on powers or capacities compatible with the truth of both forms
of determinism.

In “Freedom and Resentment,” Peter Strawson proposed another approach to
the issue of moral responsibility. Instead of turning first to an analysis of our concepts
of agency and responsibility in order to check whether these concepts were compatible
with the truth of determinism we should instead look at our actual practices of holding
others responsible and turn to an analysis of those practices for guidance. A proper

---

specification of the practice of holding responsible would therefore come to inform us about the conditions necessary for being responsible agents. Strawson's response, like his approach, was unorthodox.

Strawson's claims, explained more fully in the following section, are that responsibility ascriptions, our judgments about who is responsible, are independent of the question of determinism. The truth or falsity of determinism is otiose, according to Strawson, when it comes to the question of responsibility. This was because responsibility ascriptions turned out to depend much more on the nature of human social relations and connectedness and the role that our 'reactive attitudes' played within that network of connectedness than on metaphysical truths about determinism. Strawson's understanding of reactive attitudes and how they come to give content to our concept of a responsible agent are rehearsed below, albeit incompletely. I am ultimately looking to see if an argument supports emotional constitutiveness. Reactive attitude theorists presume that they do. Because of this, my analysis of Strawson and Strawsonians will be limited to those aspects where they come closest to offering such arguments.

6.2 What is A Reactive Attitude?

If reactive attitudes end up playing an important role in explaining what makes an agent responsible it is important to be clear about what a reactive attitude is. Depending on how those attitudes are specified the resulting view may or may not be Sentimentalist as I have been defining that term. Strawson believes that the reactive
attitudes, whatever they end up being, are components of our natural evolutionary endowment. All normal human beings will develop the capacities necessary to respond to others reactively. Reactive attitudes, for Strawson, are basic emotions. Strawson's focus is on

![5]he non-detached attitudes and reactions of people directly involved in transactions with each other; of the attitudes and reactions of offended parties and beneficiaries; of such things as gratitude, resentment, forgiveness, love, and hurt feelings.  

The first four examples are specific emotion names while 'hurt feelings' seems to be a catch all term for any negative attitude about oneself caused, perhaps, by others.  
Elsewhere Strawson speaks of “forgiveness,” “self-love or self-esteem or vanity” and “goodwill, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other” as reactive attitudes. Nearly all of these attitudes are synonymous with, or clearly connected to, emotion terms. Indifference and forgiveness are outliers in this respect though they are in the ballpark. Forgiveness is often characterized as the forswearing of resentment and indifference appears to be a cause for reactive attitudes (contempt perhaps). The point here is that reactive attitudes appear to be attitudes that presuppose social relations and which are used to participate in those relations. They also share another important feature: the proper object of the reactive attitudes are agents.

5 Strawson “Freedom and Resentment” Par. 1 Sec. 3  
6 That being the case it would seem that resentment would be a kind of hurt feeling though perhaps one especially relevant to our practices of holding responsible.  
7 Strawson “Freedom and Resentment” Par. 2 Sec. 3  
8 I borrow this characterization of forgiveness from Jeffrie Murphie's “Forgiveness and Resentment” in Jeffrie Murphie and Jean Hampton's Forgiveness and Mercy Cambridge University Press: Cambridge 1988
Love, for example, typically takes an agent (or agents) as its object and has characteristically social aspects (declarations of love) as part of its script. The same is true of contempt or forgiveness and the the rest of the attitudes on the list. For Strawson any attitude we take toward ourselves or others in virtue of their actions or social standing counts as a reactive attitude.

More recently Jay Wallace has offered a theory of moral responsibility in the reactive attitude tradition. Wallace, unlike Strawson, restricts the class of attitudes relevant to understanding our practice of holding responsible. Wallace's notion of the relevant attitudes is more exclusive than Strawson's. He argues that not all of Strawson's attitudes are reactive in the right way to function in an explanation of our practices of holding responsible.

What is this right way? For Wallace the relevant reactive attitudes include only resentment, indignation, and guilt. This is in part because Wallace makes distinctions between different attitudes we might take toward ourselves and other agents. Wallace, like Strawson, distinguishes between reactive and non-reactive (what Strawson calls 'objective') attitudes. Wallace makes a further distinction between reactive attitudes and moral reactive attitudes. Moral reactive attitudes are attitudes that respond to violations of moral obligations. Moral reactive attitudes therefore are those attitudes that respond, primarily, to violations of moral prohibitions. Moral prohibitions are composed of moral expectations that we accept and hold ourselves and others to. It is

---


10 Ibid. 34-36
because resentment, indignation and guilt are moral reactive attitudes that they are unlike the rest of Strawson's attitudes and can be distinguished as a unique class among reactive attitudes. Only these attitudes respond to violations of moral expectations that we accept.

On Wallace's account the relevant reactive attitudes are those that respond to violations of moral obligations. This does not mean that we actually feel a reactive attitude at every violation of a moral obligation, nor does it mean that lacking such a feeling implies that a violation has not occurred. Many things can affect whether or not we actually feel resentment, indignation or guilt. What matters, as we will see, for Strawson, Wallace, and other reactive attitude theorists is whether or not we believe we should respond with one of the reactive attitudes. If we think we should, if we think that the agent in question is an apt target for our attitudes then, on this family of views, that agent is responsible. The structure of this kind of view is a familiar one. It is a form of second-order Sentimentalism.

6.3 From Reactive Attitudes to Moral Responsibility

In this section I explain the role that the reactive attitudes serve for Strawson and Wallace. Although both authors believe that reactive attitudes play an essential role in explaining what it means to be a morally responsible agent, the accounts diverge when it comes to explaining why. The following section will concern itself directly with the question of constitutiveness. As in the last chapter, I canvas arguments for the emotional constitution of moral concepts. After raising what I take
to be serious worries for Strawson's view I focus on Wallace's theory because he more explicitly takes up the question of constitutiveness in the light of criticisms to Strawson's original theory.

As I suggested in the previous section, the main claim of the reactive attitude theorists is that we should look to our practices of holding ourselves and others responsible in order to understand our conception of responsible agency. Furthermore since our practices of holding responsible rely on reactive attitudes, in senses to be specified below, we must look at how reactive attitudes drive the practice.

Strawson and Wallace both attempt to show that reactive attitudes play a deep role in human social interactions. The same is true of the role that reactive attitudes play in our practices of holding others and ourselves responsible. The conditions under which we are led to think that it is appropriate to respond to an agent reactively along with those conditions under which we think it *inappropriate* to respond to an agent reactively explain what morally responsible agency consists in. The same conditions give us an analysis not only of responsible agency but an explanation of when (and why) agents are responsible for particular acts or omissions. A rough schema for the resulting two concepts is below:

**R(A):** An agent is morally responsible iff the agent is an apt target for reactive attitudes

**R(x):** An agent is morally responsible for x iff the agent is an apt target for the reactive attitudes and it is appropriate to respond to the agent reactively in virtue of his or her x-ing

Responsibility is meant to be dependent on the conditions under which we think
agents, in general, are apt candidates for reactive attitudes. If I do not think you are the kind of thing to which I should react to with resentment or love or indignation then whatever it is you are, you are *not* the kind thing that is amenable to responsibility ascriptions. Likewise, my thinking it appropriate to feel resentment at you for not greeting me at the party last night implies not only that you have the kind of agency required for responsibility but that you are actually responsible *for* the snub last night.

Strawson and Wallace both think that discovery of the conditions under which we think an agent is an apt candidate for the reactive attitudes and the conditions under which we think individual reactive attitudes are appropriate can only be discerned by analyzing the attitudes themselves. Furthermore both authors believe that this analysis explains other relevantly related moral concepts. Someone is exempt from moral responsibility when they lack the right kind of agency. Mental impairments, for example, may be enough to render a person permanently exempt from responsibility ascriptions. What makes mental impairment exempting, on these views, is the fact that we no longer think it apt to respond reactively to persons with mental impairments. In other words the claim is that it is a fact about our reactive attitudes (that we would suspend them) that unites exempting impairments into a class. Although particular impairments may have other considerations that can be marshaled in their defense, the reactive attitude theorist is committed to the claim that the only unifying feature of our exempting practices is that we think it appropriate to suspend the reactive attitudes to exempt agents.
Excusing conditions are also explained by this paradigm. Something becomes an excuse if we believe it would be appropriate to modify or suspend our reactive attitude in a case where the excuse applies. If I come to learn that the reason you did not say hello to me at the party last night was because you were speaking with one of my dreaded rivals at the time and you wanted to spare me a confrontation then I may come to think that you not only do not deserve my resentment but that you may in fact deserve praise. As with exemptions, specific excusing conditions vary greatly. What unites the class of excuses as excuses is their relationship with our reactive attitudes. Were it not for the fact that we are apt to see excuses as those conditions where we believe it merited to suspend our reactive attitudes then whatever else the condition may be, it is not an excuse.

If, as both Strawson and Wallace ultimately argue, the conditions that render reactive attitudes apt and appropriate end up being grounded on the content of our attitudes and if those attitudes are themselves not grounded on emotion-independent factors then moral responsibility can be said to be constituted by the reactive attitudes. Reactive attitude theories of moral responsibility are useful test cases for second-order Sentimentalism. Instead of having to make the global argument that concepts or values are constituted by emotional responses, reactive attitude theorists need only make the more limited argument that our concepts of responsible agency and responsibility are constituted by the reactive attitudes. In the next section I focus on what I take to be three lines of argument meant to support the constitutive claim.
6.4 Emotional Constitution

Reactive attitude theorists accept that the constitutive claim holds between core concepts in moral responsibility and the reactive attitudes. In this section I canvass what I see as three arguments that have been taken to support reactive attitude theories. In particular I examine what I take to be Strawson's basic emotions approach to constitution, Wallace's arguments for 'deep' responsibility, and finally attempts to use psychopathy as a kind of proof for the emotional constitution of our concept of responsibility. If any of these strategies bear fruit then Sentimentalism can be vindicated about at least one set of related moral concepts. Some of these argument styles are familiar, others are interestingly new but all of them, I argue, run aground against problems first introduced in the previous chapter. Although many of these arguments support motivational and epistemological roles for reactive attitudes they fall short of making the constitutive thesis the best or only explanation for our concepts of responsible agency and moral responsibility. As I argue in the eighth section of the chapter, the proceeding arguments can help us see what shape an argument for constitution might need to take.

6.5 Strawson's Arguments

Strawson, as I see him, gestures in a couple of directions regarding the reactive attitudes. The first can be gleaned from the following pair of quotations:

The human commitment to participation in ordinary inter-personal relationships is, I think, too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might so change our world that, in it, there were no longer any such things as
inter-personal relationships as we normally understand them; and being involved in inter-personal relationships as we normally understand them precisely is being exposed to the range of reactive attitudes and feelings that is in question.\(^{11}\)

And

\[\text{To change our world that in it all these attitudes were wholly suspended, I must answer, as before, that one who presses this question has wholly failed to grasp the import of the preceding answer, the nature of the human commitment that is here involved: it is useless to ask whether it would not be rational for us to do what it is not in our nature to (be able to) do.}\(^{12}\)

Strawson here invokes the 'naturalness' of the reactive attitudes, their incorrigibility and ineliminability, in order to support the view that our reactive attitudes should serve as important facts about us. Try as we might we cannot rid ourselves of these attitudes. Furthermore these attitudes are important because they make human commitment possible. Human commitments, the entire range of human inter-personal relationships, just are the practices that support the expression of the specific reactive attitudes.

Since we care so much about these commitments we cannot rid ourselves of them.

This sounds like the kind of appeal to basic emotions we have seen before.\(^{13}\) Strawson

\(^{11}\) Strawson Sec 4

\(^{12}\) Ibid. Sec 6. There is an important question here about how best to interpret Strawson's argument. We could imagine Strawson as committing himself to the ineliminability of reactive attitudes at at least three different levels. The first level would be that any particular reactive attitude type is ineliminable. For example we might claim that it is impossible to rid ourselves of the emotion of resentment. Another claim would be to say that our current suite of reactive attitudes is ineliminable such that our current practices and human commitments, as circumscribed by our current reactive attitudes are basic and therefore ineliminable. Lastly we could see Strawson as making the far more general claim that some set of human commitments must exist and that they are constituted by some set of reactive attitudes. This would make Strawson's view compatible with a wide range of constructivist positions. I think Strawson is committed to some versions of the first two interpretations. That is, Strawson seems to argue both that all of our particular reactive attitudes are ineliminable \textit{and} that the current set of human commitments we have are ineliminable and constitutive of those commitments. This is in part, in both cases, because of how Strawson understands what it means for something to be a reactive attitude.

\(^{13}\) Michael McKenna has called this argument the 'Psychological Impossibility Argument.' Although
commits himself here to various claims: (1) Reactive attitudes are natural (a part of our evolutionary heritage) (2) Reactive attitudes ground human commitments (3) Reactive attitudes are ineliminable and (4) Responsible agents are agents capable of making commitments.

The argument is that all of the facets of moral responsibility can be traced back to the norms that govern the application of reactive attitudes. Furthermore because these attitudes are 'natural' they are taken to be universal (or nearly so). Strawson seems to argue for the constitutive role because of its explanatory power. The conditions under which we think it is appropriate to hold someone responsible are (entirely) explained in terms of the conditions under which the reactive attitudes are appropriate. The forms of responsibility we care about can be fully explained by reference to the reactive attitudes. An excuse is an excuse because of the way our attitudes are modified in their presence. The argument then runs that because those attitudes are natural they are the best explanation for the universality of our practices and hence that they are the best explanation for the content of the concepts. A suggestion in that direction can be gathered out of something Strawson says later still when considering whether we would not be better off ridding ourselves of the reactive attitudes:

McKenna thinks Strawson does derive a key insight about moral responsibility in “Freedom and Resentment” he does not think the Psychological Impossibility Argument is the place to find it. Instead he argues that a closer look at the argument from excuses will show us that one of the unifying features of appropriate reactive attitudes is that they focus on the quality of an agents will. This is likely true but for my purposes this points more toward an explanation of how we locate the locus of concern for a reactive attitude. See McKenna's “Where Frankfurt and Strawson Meet” in Midwest Studies in Philosophy 29 (1) 2005: pp. 163-180.
Such a question could seem real only to one who had utterly failed to grasp the purport of the preceding answer, the fact of our natural human commitment to ordinary inter-personal attitudes. This commitment is part of the general framework of human life, not something that can come up for review as particular cases can come up for review within this general framework.\footnote{Strawson Sec 4}

Paul Russell has criticized Strawson's approach here by arguing that there is an inherent tension between the claims that our reactive attitudes are ineliminable and the arguments meant to support the attitudes themselves as involving fundamental human commitments.\footnote{See Paul Russel's “Strawson's Way of Naturalizing Responsibility” \textit{Ethics} 102 (2) 1992: p 291-292} The tension is that the claim that the reactive attitudes are ineliminable renders the justificatory arguments useless because to presuppose that the attitudes require justification is to suppose that they can be modified but the ineliminability claim makes clear that we are stuck with the reactive attitudes. Russell argues that Strawson conflates two senses in which we could hold a naturalist picture of reactive attitudes: type and token naturalism. The details of the distinctions between these two naturalisms and indeed of Russell's critique of Strawson are not as important for my purposes here as his critique against Strawson's assumptions that the naturalness of a reactive attitude necessarily implies our inability to modify or entirely refrain from using the reactive attitudes.\footnote{For the interested, Russell argues that type naturalists see ourselves as unable to rid ourselves of our reactive attitudes. Token naturalists argue that it is impossible for us to see every instance of our reactive attitudes as inappropriate. Russell argues that it matters, crucially, which of these two positions Strawson intends to be taking.} I think here Russell is on to something. The mere fact of naturalness, for example, does not imply that we can not modify our fear responses in particular cases. Excuses are testament to that. Something more than mere
ineliminability or naturalness must therefore be involved.

It seems entirely possible for us to come to learn something about either the objects of our fears or about ourselves that would render fear *always* inappropriate.\(^{17}\) Perhaps, as I suggested in the previous chapter, humans, through the use of sufficiently advanced biotechnology, become essentially immortal beings. Our capacities for healing are such that nearly all injuries are survivable. In such a future many kinds of fear would cease to be appropriate just as fear of ghosts has ceased to be appropriate for most of us. In this kind of world we would rid ourselves of fear not because we manage to literally extinguish the emotion but because we come to the conclusion that it is never appropriate. On the rare occasions when we feel afraid we might come to enjoy the experience (as we do with thrill rides) or quickly extinguish it because it is inappropriate. Similarly a post scarcity society, one where energy is so abundant that goods and services are essentially free, may be one in which all, or nearly all, forms of envy would be inappropriate. Envying your neighbor's new car or house becomes irrelevant when you can also effortlessly acquire the same car or the same kind of

---

\(^{17}\) Gary Watson offers a similar critique of Strawson's position here. Although we may not be able to rid ourselves of the reactive attitudes it may turn out that the attitudes themselves are sensitive to determinism in such a way that the truth of determinism would render our attitudes inappropriate. Further, if membership in a moral community requires sharing the values of that community the, Watson argues, it seems as if the most evil agents among us are, in virtue of their evil outlook, not a part of our moral community and hence are excused for their behaviors. Both conclusions are problematic. See Watson's "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in F.D. Schoeman (ed.), *Responsibility, Character and the Emotions: New Essays in Moral Psychology* Cambridge: Cambridge University Press, 1987: pp. 256-286; though see also Michael McKenna's "The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism" in *The Journal of Ethics*, 2 (2), 1998: pp. 123-142 for a response to Watson's criticism. McKenna argues that being a member of the moral community does require both an understanding of moral address and a shared set of goals that that address satisfies. McKenna argues that although evil agents may thus fall outside the moral community, so long as they have the *capacity* to be a part of the moral community then they may be held accountable by it. This capacity can be demonstrated by an agent that knowingly flouts those values.
house. Strawson is therefore concluding too much when he moves from the naturalness of the reactive attitudes to something in the neighborhood of the constitutive claim.\textsuperscript{18}

But the second line of argument also bears mentioning, if only briefly. Strawson has claimed that human relationships cannot exist without the reactive attitudes and that the practice of holding responsible is one such commitment. There are two assumptions necessary to make this argument work. The first is that it is true that human commitments are dependent upon reactive attitudes. The second is that this dependency is best explained by seeing the relation between commitment and attitude as constitutive. It seems right to question the basic emotions approach Strawson is invoking here. The list of reactive attitudes that Strawson invokes do not appear on most lists of basic emotions created by those researchers who genuinely believe that they are out there.\textsuperscript{19} This is problematic since it creates an empirical hurdle for the Strawsonian that she is unlikely to be able to meet. Furthermore there is the added risk of committing oneself to the existence of basic emotions generally. They are not likely to be out there at all. The loss of basic emotions renders Strawson's universality claims question begging (much as it did the universality claims of the Rational

\textsuperscript{18} There is another, further, critique here suggested by Jay Wallace. According to Wallace, Strawson is here uncharitably casting the incompatibilist as offering an 'external' critique against the reactive attitudes: that we should remove them because they are incompatible with determinism. Understood in this external way Strawson's argument about the ineliminability of our reactive attitudes hits home. Wallace suggests however that the incompatibilists are offering an internal critique: if determinism is true then our reactive attitudes are always inappropriate. The simple ineliminability of our attitudes wouldn't extinguish this critique. See \textit{Responsibility and the Moral Sentiments} pp. 97-98.

\textsuperscript{19} Love and forgiveness are nowhere to be found in Ekman's standard list of basic emotions for example. See Paul Ekman's “Basic Emotions” in \textit{The Handbook of Cognition and Emotion}. T. Dalgliesh and T. Power (eds) U.K. John Wiley and Sons. 1999
Sentimentalists). But beyond these empirical worries we can question whether the assumptions above really get Strawson the strong constitutive claim he seems to want. We could cash out dependency claims between reactive attitudes and responsible agency along any of the familiar roles for emotion laid out in previous chapters. Reactive attitudes could be playing motivational, epistemological, or constitutive roles in our practices of holding responsible. The truth of any of these roles would make the first assumption true (that human commitments are dependent on reactive attitudes). Because of this, the second assumption does not follow from the first. Strawson clearly wanted constitution here in order to sidestep the debate between compatibilists and incompatibilists but his arguments do not get him this far.

Jay Wallace does better, I argue, by dropping the commitment to basic emotions and seeking to show, via a constructivist route, that our conceptions of responsible agency are constituted by resentment indignation and guilt. I turn to his arguments in the next section.

6.6 Wallace and 'Deep' Responsibility

Wallace's view is subtler than Strawson's. This is not the proper place to fully articulate the nuances of Wallace's theory of moral responsibility however. My chief concern here is to explain how Wallace views moral responsibility and how reactive attitudes are involved within that theory. Wallace explicitly seeks to explain a form of response-dependence between reactive attitudes and moral responsibility. To do this Wallace turns to what he calls the 'reactive account' of moral responsibility. Wallace's
reactive account of moral responsibility defines the conditions for being a responsible agent as well as conditions for being responsible for an action in terms of the conditions under which it would be appropriate to feel resentment indignation or guilt. Wallace sees himself as advancing the Strawsonian point about the reactive attitudes but in a more philosophically rigorous way.

Wallace’s *Reactive Account of Moral Responsibility*, like Strawson's own view, should be understood as a second-order Sentimentalism:

The basic stance of holding someone morally responsible involves a susceptibility to the reactive emotions if the person breaches moral obligations that we accept, or the belief that it would be appropriate for us to feel those emotions if the person should violate those obligations.\(^\text{20}\)

An agent is morally responsible for something if we either feel an appropriate reactive attitude toward her or if we think that it would be appropriate to feel one even if we do not. The reactive attitudes are appropriate only when they satisfy what Wallace calls the two basic conditions of fairness.\(^\text{21}\) It is *unfair* to hold someone responsible if she either has not done anything wrong, that is, if she has not violated a moral norm that

---

\(^{20}\) Emphasis is mine. The disjunction here is (1) not exclusive and (2) because of 1 allows Wallace to ground his account of holding responsible on reactive emotions, in general, without requiring them agents to feel them all the time (or even most of the time). It requires knowledge of what resentment, indignation, and guilt feel like such that we can know when we should feel them but doesn’t require that we feel them at any particular time. Wallace refers to this way of specifying his view as Schema (N). See Wallace p. 66

\(^{21}\) Wallace believes that our reactive attitudes are appropriate only when two conditions of fairness are met. *The Principle of no Blameworthiness Without Fault*: Those who have not in fact done anything wrong clearly do not *deserve* to be subjected to the reactive emotions and the forms of sanctioning treatment that express them. This principle governs excusing conditions. See p. 135. A second principle, governing exempting conditions, *The Moral Principle of Reasonableness*:[“]It is unreasonable to demand that people do something—in a way that potentially exposes them to harms of moral sanction—if they lack the general power to grasp and comply with the reasons that support the demand.” See p. 161.
we accept, or if she is unable to understand or respond to moral demands. Being unable to understand or respond to moral demands is, for Wallace, bound up with being able to respond reactivity.

If we are tempted to think that we should just turn to the conditions of fairness for answers to moral responsibility then we are making, as I see it, one of two mistakes according to Wallace. The first mistake is to assume that there is a realm of facts about moral responsibility that are independent from and that explain our reactive attitudes that we can turn to in the first place. This means that we are assuming that the conditions of fairness exist prior to and independent from our reactive attitudes and we assume further that the reactive attitudes are serving merely as heuristic guides in the epistemological role. To assume the existence of this realm of reactively independent and prior facts about moral responsibility is problematic according to Wallace. This is because the set of *practices* we make use of for holding ourselves and others responsible relies upon a framework of reactive emotions and moral obligations and the sanctioning behavior that expresses those emotions. Our practices, therefore, seem fully explainable without appealing to metaphysical facts about responsibility. It seems odd, Wallace argues, that we should think that there exist facts that are prior to *and independent* from the reactive attitudes that ultimately justify them. It is also mysterious that these facts would be unnecessary to explain our practice of holding agents responsible.

The best explanation for why our conception of responsible agency has the
contour that it does, on Wallace's view, is because our reactive emotions are only appropriately applied to some actions that agents do (those agents and actions that survive the conditions of fairness). We would not posit the conditions of fairness if we did not think that they set out the conditions under which our reactive attitudes are appropriate. Had our reactive attitudes been appropriate under different conditions it appears Wallace is ready to argue that the conditions of fairness themselves would be different.

This argument however, by itself, does not go far enough. In some ways it does not get us much traction at all. Wallace has, for all I have said so far, provided us with a mere description of our moral practices. These practices are also bound up with other moral considerations we care about, considerations of fairness for example. But Wallace's attempt to shift the burden of proof onto his detractors is in itself unfair. It may be the case that the reason why our reactive attitudes are appropriate only when specific conditions relating to fairness have been met could be explained by the independent importance of fairness and the fact that we train reactive attitudes to respond to these conditions as part of the process of moral education. That is, fairness explains our practices of holding others responsible and not the reactive attitudes. This seems at least as plausible as Wallace's alternative.

A further concern about this argument is that it appears dangerously circular. If we begin with an appeal to the role of the reactive attitudes in building up our notions of fairness and then test those notions of fairness by seeing whether they agree with
our reactive attitudes then the circle becomes viciously tight. Wallace may offer a rejoinder at this point however. Wallace can claim that he is really making two different claims here, one empirical and the other explanatory, and that the circularity critique rests on mistaking the evidence for the empirical claim as the argument for the explanatory claim. The empirical claim is that there is, in fact, nothing beyond the reactive attitudes that is necessary to explain our practices of holding one another accountable. The explanatory claim is that the appropriateness of the reactive attitudes is what best explains our excusing and exempting conditions. Our appeals to fairness in discussions of responsibility ascriptions would therefore result in part from the empirical facts about appropriate reactive attitudes coupled with the explanatory inference. Because Wallace, contra Strawson, believes that the reactive attitudes are not basic or universal Wallace need not be committed to the view that our conception of fairness is entirely reactive. He need only claim that the notions of fairness germane to ascriptions of responsible agency or responsibility are unified, and therefore best explained, by their dependence on appropriate reactive responses. The argument is meant, as I understand it, to shift the burden onto the detractor to find a non-reactive unifying feature for the conditions of fairness germane to responsibility. Of course most of Wallace's opponents think they have such an account. The problem becomes clearer, and more severe, for Wallace if these conditions for fairness can indeed be given a plausible response-independent reading.

This is not a trivial problem as Wallace himself does come close to admitting
the independent importance of conditions of fairness:

[M]oral norms of fairness have a privileged position in determining what it is to be a responsible agent: they set the standards of appropriateness in terms of which schema (N) is to be interpreted.  

One important feature of Wallace's account is that it is disjunctive. We need not feel resentment, indignation or guilt on any particular occasion in order to judge that someone (including ourselves) is an apt target for the reactive attitudes. This disjunctive aspect brings with it a ready objection that gets at the heart of Wallace's commitment to the constitutive relationship between reactive attitudes and our concept of moral responsibility. If it is not really necessary to be subject to reactive emotions for one’s stance to count as the stance of holding someone morally responsible, why bring in such emotions at all? Could we not understand moral responsibility just as well in terms of the fair acceptance of moral obligations, and beliefs that such obligations have been violated by agents who can understand moral demands?

Wallace's response to this worry makes his commitment to the constitutive role clear:

This [kind of response] fails to capture the deep assessment of blame and would turn it into a way of describing what an agent has done instead of reflecting an attitude towards the fact that the agent has acted wrongly. The reactive emotions are needed to capture this deep assessment and its natural connection with sanctioning behavior. If the reactive emotions are needed in order to account for moral blame then the reactive emotions are needed in order to account for blameworthiness.  

This response requires careful analysis. What does Wallace mean by a 'deep'

---

22 See Wallace, 94. Schema (N), recall, refers to Wallace's disjunctive analysis of moral responsibility in terms of either experiencing an occurrent instance of resentment, indignation, or guilt, or judging that it would be appropriate for you to feel these emotions in some particular instance.

23 Wallace 78
assessment of blame? In the passage above it appears that merely describing or evaluating what an agent has done is not deep in the right sense. A deep assessment, I take it, is deep because it connects an explanation of the relationships between agents with an explanation of our practices of holding agents responsible. Furthermore deep assessments help explain the urgency of our practices. To judge that someone is responsible is to demand action and reactive attitudes help explain why these different thoughts, judgments, motivations, and practices are united together. But this does not go far enough. It remains open to Wallace's detractors to argue that our reactive attitudes do indeed make us care about our responsibility ascriptions (in the sense that they give them urgency, that they motivate us) and they may even factor into a full explanation of our retributive practices of holding accountable. Still, this does not show that the reactive attitudes are constituting the concepts. The critic could here accept that Wallace's arguments support the view that the reactive attitudes frequently play the motivational role and may, via appropriate reactive attitudes, be open to accepting that reactive attitudes can also play an epistemological role. It does not defuse the complaint that independent conditions of fairness explain the concept at least as well even if reactive attitudes are necessary to explain our practices.

Although Wallace never explicitly links this worry with another strand of argument in *Responsibility and the Moral Sentiments*, an interesting suggestion for responding to this worry exists. I mention it briefly here and discuss it further in section eight.
Wallace pauses in several areas to consider the application of his view to cultures which are not susceptible to resentment, indignation, or guilt. He calls these cultures “shame-cultures” because they make use of shame to enforce norms for social cooperation. Since Wallace denies that emotions are basic it becomes a live option whether human societies end up holding one another accountable using resentment, indignation, and guilt. Although little anthropological or sociological data is produced to support this point, I believe a fuller examination of the reactive emotions different cultures use to govern cooperative norms may prove instructive. In particular it could motivate the claim that reactive attitudes constitute what it means to be a responsible agent if the concepts and practices of some cultures diverge sufficiently from our own.

Whatever motivation is produced by this kind of argument would certainly be augmented by more precise arguments meant to directly support the constitutive claim. In the next section I explain, and criticize, arguments that aim to show that psychopathy supports the constitutive claim.

6.7 The Allure of Psychopathy

Psychopathy is a controversial condition. It does not appear in the latest addition of the *Diagnostic and Statistical Manual* used by psychologists to diagnose patients though it does have a set of well known diagnostic criteria developed by Robert Hare. In what follows I explain how psychologists and neuroscientists

---

24 The standard measure of Psychopathy is the Hare Checklist (Revised). Most research on psychopathy uses this diagnostic tool to identify psychopathic agents. There are at least two reasons why the DSM doesn't include psychopathy. The DSM focuses on observable behavior criteria since it is a diagnostic manual. A diagnosis of psychopathy requires inferring emotional/mental capacities that may not have obvious behavioral signs. Furthermore psychopathy lacks treatment options. There is reason for differentiating between ASPD and Psychopathy however. ASPD is a much more
understand psychopathy and then explain why philosophers have been attracted to this condition as a test case for the relationship between reactive attitudes, responsible agency, and ascriptions of moral responsibility. I find promising directions for an argument that motivates the constitutive claim. This argument is sketched out in the eighth section of the chapter.

Psychopaths are noted for being charming, manipulative, and consistent liars. Although not entirely devoid of affect psychopaths are noted for living a parasitic lifestyle, and for having, usually, a history of criminal behavior.\textsuperscript{25} They are also unique in their readiness to use instrumental aggression to achieve their ends. Instrumental aggression is contrasted with reactive aggression. Reactive aggression is the kind of aggression we see as uncontrolled outbursts or fits. Instrumental aggression is the result of a choice to use force to achieve one's ends. Psychopaths are also known for having a difficult time distinguishing between moral violations, like murder, and violations of convention (e.g. slurping one's food in a restaurant).\textsuperscript{26} This list of behavioral traits can be augmented by a more precise account of the aberrant

\begin{footnotesize}
\begin{itemize}
\item Psychopathy, it's important to note, comes in degrees. While most of us would rate quite low on Hare's psychopathy checklist, many individuals would rate as falling below the diagnostic threshold for psychopathy but far above the average. See Robert Hare and Craig Neumann, “Psychopathy” in \textit{Oxford Textbook of Psychopathology} 2nd ed.. Oxford University Press New York, NY. p. 622-650
\item See Mairead Dolan and Rachael Fullam“Moral/conventional transgression distinction and psychopathy in conduct disordered adolescent offenders.” \textit{Personality and Individual Differences} 49, 2010. p. 995–1000. Some philosophers have argued that the moral/conventional distinction does not track what it appears to track. Virtue theoretic notions of morality, for example, don't map on to the distinction neatly and that therefore the question of whether psychopath's divergence on the moral/conventional task is relevant to the question of their responsibility becomes dependent on one's underlining meta-ethical position. See Manual Vargas and Shaun Nichols “Psychopaths and Moral Knowledge” in \textit{Philosophy, Psychiatry, & Psychology} 14, 2, (2007). pp. 157-162
\end{itemize}
\end{footnotesize}
capacities of the psychopath. While most of us exhibit these traits on some occasions what makes psychopaths different is that they regularly express these traits and do so for very different reasons than when we, on occasion, behave in similar ways. It is because psychopaths exhibit aberrant capacities for understanding empathy, agency salience, and aversive conditioning that they present with the behaviors above. These behaviors are strongly indicative of the cognitive and affective differences thought to exist between psychopathic and non-psychopathic agents. Reactive attitude theorists argue that if these differences can be shown to plausibly change the reactive attitudes psychopaths have or affect the ways in which psychopathic agents understand their own reactive attitudes then this should be taken as support for the claim that the reactive attitudes are playing an important role in our conception of moral responsibility.

We should begin with empathy. Psychopaths have trouble with empathy. Psychopaths do not respond to the emotions of others as normal agents do. When presented with emotional cues (sad faces, recordings of crying, etc) most agents show physiological and behavioral analogues of the emotions they are exposed to. Psychopaths do not tend to respond in this way. On most measures of empathic response psychopaths remain flat. Let's call this deficit a lack of empathic distress.

It is also harder for psychopaths to understand the relationship between their actions and blame. Normal agents automatically see the consequences of their actions

---

both as resulting from their agency and as attributable to them. Psychopaths find it harder to see their actions in this way. They are more likely to attribute their actions to something other than their own agency, they make 'external' blame attributions (they blame their situation or society) instead of their own agency and judge themselves less responsible (in light of these external factors) for the consequences of their actions. Call this deficit a lack of agency salience.\textsuperscript{28}

Last, for my purposes, psychopaths also have trouble with aversive conditioning. Aversive conditioning is the ability to learn associations as a result of pairing an object or image with an unpleasant stimulus. So, for example, pairing an image of the countryside with a foul odor would lead, in normal agents, to characteristic skin conductance response changes associated with disgust when later shown the image of the countryside alone.\textsuperscript{29} Psychopaths do not seem to have this capacity. Instead they tend to have flat or very diminished reactions to aversive conditioning. Important to note here is that aversive conditioning is an important component for much of our practices of moral socialization. Being punished for behaving badly is one form of aversive conditioning. Feeling guilty at acting wrongly is another, internal, form of aversive conditioning. Both of these methods of moral socialization are degraded in the psychopath.

These aberrant capacities help explain the set of psychopathic behavioral

\textsuperscript{28} Studies of blame judgments in psychopaths are difficult to come by and, like most studies involving psychopaths, are composed of a small number of participants. For one study on blame attributions in female psychopaths see Weizmann-Henelius et al. “Violent Women, Blame Attribution, Crime, and Personality” in Psychopathology (2002) 35(6):355-61.

\textsuperscript{29} See Herta Flor et al 2002: pp. 505-518.
characteristics. Their lack of empathic distress means they do not see the pain of others as painful. Their trouble with agency salience means they have a much harder time seeing the results of their actions as stemming from their deliberative processes. Finally psychopaths have a demonstrated difficulty with aversive conditioning. This explains why psychopaths fail to see the suffering they cause to others as relevant and why they fail to learn to modify their behavior in light of the suffering they cause to others. A life characterized by lying, parasitism, instrumental aggression, and criminal and anti-social behavior makes sense given those traits.

Given this characterization of psychopathy it should be clearer why they make for good test cases regarding the relationship between reactive attitudes and responsible agency. If it is necessary, as it is on Wallace's view and on many others, that an agent be able to understand moral reasons in order to be morally responsible then the psychopath's affective deficits may impact this capacity. As I have shown, the stunted affective capacities of the psychopath affect the way in which psychopaths reason and the way in which they understand moral demands. This implies at least some form of the epistemological role for the reactive attitudes. The question then is whether these stunted capacities are so stunted as to render psychopaths non-responsible. But reactive attitude theorists want to claim that psychopaths can do more than support even a strong epistemological role. Although I think psychopathy by itself will not take us further than an epistemological role for the reactive attitudes I do think they fit into a larger argument for the constitutive role. I ultimately argue that the
psychopath's deficits, because they are primarily affective lends prima facie credence to the claim that these affective capacities are necessary in a strong sense that, when combined with cross-cultural data on shame-cultures, can be used to construct a prima facie case for the constitutive claim. For the moment however I have yet to say how psychopaths have been thought to support reactive attitude theories.

Jesse Prinz has at times proposed to use psychopaths for just this purpose. For example, he has asked us to

- Imagine a person who knows everything non-emotional about killing. She knows that killing diminishes utility and that killing would be practically irrational if we universalized the maxim, thou shalt kill. Would we say of this person that she believes killing is wrong? It seems not. She could believe all these things without having any view about the morality of killing or even any comprehension of what it would mean to say that killing is wrong. Conversely, if a person did harbor a strong negative sentiment towards killing, we would say that she believes killing to be morally wrong, even if she did not have any explicit belief about whether killing diminished utility or led to contradictions in the will. These intuitions suggest that emotions are both necessary and sufficient for moral judgment.

Here Prinz argues that emotions are necessary for moral judgments. It might seem as if descriptive claims about a state of affairs do not count as moral, on his account, without the aid of a further emotional component. In fact he says something quite stronger than this. Moral judgments require only the emotional component given the intuition Prinz is pressing above. The argument depends on our intuitive judgments about judgments here. Prinz is betting that we would infer that an agent is making a genuine moral judgment if her judgments are made reactivity. Psychopaths, for many,
are the kinds of agents that push this intuition forward.\textsuperscript{31}

Prinz elsewhere argues that psychopaths may help to lend support to the metaphorical component of a view he calls strong emotionism. On my terms his strong emotionism can be described as a commitment to the constitutive role. If moral facts were not essentially related to emotions, his argument goes, then there ought to be other avenues that allow a psychopath access to moral truths in a non-parasitic way. That is, psychopaths, being rational agents, should be able to come to have moral knowledge without depending on their emotionally 'normal' neighbors. But psychopaths have trouble understanding the difference between moral violations and violations of convention. This suggests that they do not actually understand what it means for something to be a distinctly moral judgment. This is grist for Prinz's mill since it suggests that this deficit stems from the psychopath's affective deficits. It is, after all, in their interest to appear normal and a passing understanding of morality would go a long way toward blending in with, and taking advantage of, other members of society. The fact that psychopaths cannot do this provides, according to Prinz, some plausibility to the claim that moral concepts themselves are constituted by emotions.\textsuperscript{32}

\textsuperscript{31} Shaun Nichols has performed informal studies that appear to contradict the assumptions that Prinz is appealing to here about our judgments of psychopathic moral knowledge. Nichols' results suggested that at least undergraduates disproportionately rate a psychopathic agent who acts badly on the basis of self interest (while also claiming to know that self interest is not universalizable) as having an actual understanding of and ability to make moral demands. Interestingly they also judged that a psychopathic agent who acts virtuously on the basis of its rationality ('I do the virtuous thing because it is rational') does not display a genuine understanding of morality or moral demands. Although Nichols thinks this suggests that the platitudes surrounding moral judgments show that they are not necessarily rational and therefore that Michael Smith's famous analysis of moral rationalism is incorrect, Nichols' data also appears to conflict with Prinz's intuitions as well. This would lessen the force of Prinz's intuition pump. See Nichols' "How Psychopaths Threaten Moral Rationalism, or Is it Irrational to Be Amoral?" in The Monist (85) 2002.

\textsuperscript{32} Jesse Prinz The Emotional Construction of Morals p. 46
Paul Russell is attracted to a similar claim. He too thinks that psychopaths lack a uniquely moral standpoint and that this lack can be traced to their affective deficits:

What [psychopaths] cannot do, however, is accept these reactions and sanctions in the same way as a normal person can. Whereas the person with moral sense is capable of accepting these responses in a way that involves coming to feel them from the inside or sharing these negative sentiments, this is simply not possible for the psychopath ...he cannot, therefore, experience the grip and force of moral considerations through the channel of reactive sentiments themselves. Without an ability of this kind an agent’s ability to recognize the salience and significance of moral claims, and to be motivated effectively by them, will be radically impaired. 33

These argument are in some ways hard to assess. Insofar as psychopath’s lack an 'urgency' to their moral beliefs this implicates the motivational but not the constitutive role and seems bound up within larger debates between motivational internalists and externalists. The same implications seem to lie behind the appeals to an 'internal' understanding of moral claims. As I argued in chapter three, internalism, the motivational role for emotions, is often a commitment that Sentimentalists find especially attractive. But it is a mistake to see the ability of Sentimentalism tp explain the internalist intuition via the motivational role as evidence for something more, the constitutive role.

Russell also suggests that psychopaths lack access to moral concepts or to a unique kind of moral standpoint. Further this lack of access or perspective can be explained by a lack of specific affective capacities. Insofar as this is the nature of the argument then this seems to support the epistemic and motivational roles for emotion.

33 Paul Russell “Responsibility and the Condition of the Moral Sense” in Philosophical Topics 32 (1-2), 2004: pp.15
but not the constitutive one unless the moral standpoint simply consists in the standpoint of reactive expression. There are innumerable mind-independent facts which humans can only have second hand information about (most scientific facts function this way) so, by itself, this argument will not take us very far. It would be circular to claim that the missing standpoint is an emotional standpoint. Russell and perhaps Prinz want to say that the affective capacities missing in the psychopath, what I called a lack of empathic distress, agency salience, and a lack of aversive conditioning, are necessary in order to enter the moral perspective. The best explanation of these appeals is that the moral perspective is one where making moral judgments involves a susceptibility to feel and respond with the reactive attitudes. This makes sense of Prinz's intuitions that simply evincing a reactive attitude is, by itself, enough to render a judgment a moral judgment. Furthermore this intuition appears more widely shared though still bound up in larger debates about the nature of moral judgments and internalism.\(^{34}\) One of Strawson's key insights was to ask questions about the proper stance from which we can make and assess judgments about moral responsibility. Insofar as this stance is bound up with reactive attitudes the intuition that attitudes like resentment, indignation, and especially guilt are ineliminable features of this stance can help form part of a larger argument to motivate the

\(^{34}\) John Martin Fischer and Mark Ravizza seem to make a very similar claim when they discuss psychopathy. Although they agree that psychopaths are able to make judgments with what appear to be moral content, i.e. 'that other people's interests have to be suitably balanced with my own,' they argue that this does not demonstrate that psychopaths have “appropriate receptivity” to moral reasons and hence that they do not count as being able to understand moral reasons. I argue in unpublished work that this implies that Fischer and Ravizza are, despite their claims to the contrary, tacit motivational internalists. See Fischer and Ravizza's, Responsibility and Control: A Theory of Moral Responsibility. Cambridge: Cambridge University Press 1998: especially pp. 79.
Something is different about psychopathic agents. We can identify statistical behavioral and neurological differences about them and from those infer something about their capacities. The psychopath's spotty performance on identifying the moral/conventional distinction, a distinction about the severity and authority-dependence of norms, coupled with their ready use of instrumental aggression and lack of empathic distress imply a disconnect from standard moral norms. Russell draws this as an internal/external form of knowledge while Prinz and others prefer to say that psychopaths are not making moral judgments at all. While these issues can be characterized as epistemological faults the psychopath may have than mere epistemological faults. Although psychopathy by itself will not secure constitutiveness it can play a role. In the following section I suggest a way of incorporating the data on psychopathy as part of an empirical test for the constitutiveness of the reactive attitudes.

6.8 Arguments on the Horizon

Wallace, I noted, mentions cross-cultural comparisons of norm enforcing practices to motivate the constitutive claim. This appeal can be used to supplement Prinz's argument. The goal here is to motivate us to question whether the fault of the psychopath is merely epistemological. If truths about responsibility and responsible agency were independent from the reactive attitudes then we should expect most normal agents to come to have access to roughly similar concepts about agency and
responsible agency in much the same way as we expect some convergence on other more obvious mind-independent facts like physical facts. If this is not the case, if agents with normal affective capacities come to have not only differing reactive attitudes but also differing understandings of when it is appropriate to enforce norms at all, then it seems possible to motivate the claim that there is something more than the epistemological role at work. Psychopathic agents show us that if reactive attitudes are playing a merely epistemological role then it is at least a very strong form of the epistemological role. Only reactive attitudes give us an understanding of morally responsible agency and the conditions for accountability. However if non-psychopathic agents shared similar conceptions of responsible agency, especially if their reactive attitudes were different, then we should be motivated to think that reactive attitudes are at best serving *only* an epistemological role and that different cultures use different reactive attitudes to arrive at the same set of response-independent facts about responsibility. The existence of a set of facts about responsible agency that exist independent of the active attitudes would be the best explanation were this to be true. But this is not the case and that is a significant empirical finding.

Psychopathic conceptions of responsibility diverge dramatically from our own despite their having *some* affective capacities and it being in their self-interest to acquire our concepts. Furthermore if Jay Wallace, Bernard Williams and others are right then the conceptions of accountability and accountable agency that exist in cultures whose pro-social norms are governed by different reactive attitudes diverge
from our guilt-based responsibility attributability ascriptions. This is hard to explain if the reactive attitudes were solely playing the epistemological role but fits more comfortably within a view that has the reactive attitudes playing a constitutive role.

Prinz argued that psychopaths, being otherwise rational, should be motivated to acquire our moral concepts. Psychopaths, however, do not seem to acquire our moral concepts. What they do seem to acquire are concepts about how others behave when they behave morally and what other agents expect of one another. This is Russell's *external* sense of having a moral concept. This claim, by itself, did not motivate the constitutive role for emotion without begging the question. For example, in order for the psychopathic method of concept acquisition to fail as counting as actually having a concept one would need to already believe that having a moral concept necessarily requires affective or motivational input. This begs the question against the motivational externalist who would be open to interpreting the psychopath as an amoralist: an agent who understands but rejects—and is therefore not motivated by—moral requirements. Psychopathic behavior does seem to imply that psychopaths either do not actually understand moral concepts or, if they do, that they have significant difficulty translating their moral concepts into moral actions.

When we couple this fact about psychopath's errant abilities with Wallace's suggestion that other cultures could potentially enforce norms of social cooperation using a set of reactive attitudes other than resentment, indignation, and guilt we get an interesting empirical test of the reactive attitudes (ETRA), the results of which may
work in favor of those advancing the constitutive claim:

**ETRA:** If the concepts of responsibility and responsible agency were independent of reactive attitudes we should expect agents, normal or aberrant, without the relevant attitudes to have access to those concepts. If we consistently fail to find agents with these concepts or find agents with diverging concepts then the concepts should be seen as likely candidates for being response-dependent.\(^{35}\)

Why should we think that ETRA is plausible? It would seem odd if otherwise normal agents, indeed sets of otherwise normal cultures, did not make use of concepts of responsible agency or responsibility if these concepts are response-independent moral concepts. Since those concepts appear to determine who should be held accountable and under what circumstances, they stand to greatly enrich cultural practices were they to exist in this response-independent way. Psychopathic agents and Wallace's shame-cultures help fill in the rest of the intuitive pull of ETRA here. We already saw that Prinz argues that it is in the interest of any psychopaths to acquire our moral concepts. Their poor ability to explain the differences between moral violations and violations of convention suggest that their understanding of the distinction is poor. If, within our own culture, psychopathic agents fail to grasp the concepts of attributable agency or accountability and this is because of their aberrant capacities for empathic distress, agency salience, and aversive conditioning then this is evidence for the reactive attitudes playing the epistemological role; further, if other cultures do indeed base the enforcement of norms of social cooperation on attitudes very different from

---

\(^{35}\) A further stipulation here is that humans wish to align their practices with truth. Insofar as this assumption is true we should expect agents to seek out these response-independent facts about moral responsibility.
resentment, indignation and guilt and these cultures also make use of different evaluative concepts to govern the practices that enforce social cooperation then it seems plausible to think that the relation between reactive attitudes and the content of the concepts of responsible agency or responsibility is stronger than an epistemological relation. Why should we think this?

If we can indeed find a candidate culture for ETRA I argue that the best explanation for the differences between a candidate culture's practices and norms and our own is that the differences in norms is explained by the differences in reactive attitudes and norms developed to govern those attitudes. The other options on the table, for example that cultures are making mistakes or that they are more primitive than our own, require at least as much further motivation as the claim that the evaluative concepts are constituted by the reactive attitudes. But ETRA is not meant to prove, once and for all, that the constitutive thesis holds between an evaluative concept and a subjective human response. I see ETRA as giving plausibility to the constitutive role but also as shifting the burden onto its detractors to better explain these stable divergences without appealing to claims that a culture is more primitive.  

Bernard Williams' analysis of classical Greek moral concepts potentially points us to a candidate for ETRA. Williams at times seems to argue along the same lines in

---

36 Pluralism might seem like the proper response here from the response-independent opponent. It is open to them to claim that there are a great many real values out there. If those values are incommensurable then we should indeed expect different groups to appeal to some sets of commensurable mind-independent values while others latch on to other sets. I think this is a sensible move on the part of the response-independent theorist though it brings with it the explanatory burden of justifying the pluralistic picture under description. I would claim that here too response-dependent analyses are at least as able to explain this pluralism as the response-independent analyses.
his *Shame and Necessity*. He is there concerned with identifying classical Greek notions of agency and responsibility and their connections with reactive attitudes like shame and agential concepts like autonomy. One preliminary conclusion he reaches is friendly to the ideas in this section:

> What I do deny is that such emphasis as we lay on the voluntary is supported at some deeper level by a basic idea of what it is to be “really” responsible...[w]e deceive ourselves if we suppose that public practices of ascribing responsibility can be derived from an antecedent notion of moral responsibility, or that the idea of the voluntary is uniquely important to responsibility.

In line with the reactive attitude theorists, Williams here argues that it is a mistake to believe that an analysis of responsibility should begin with an analysis of the concept of responsibility. The public practices must first be understood on their own terms. Shame appears a central emotion within classical Greek culture for Williams. Williams draws an important divide between shame and guilt in terms of the attitudes that each elicits and these go on to play an important role in his characterization of the development of morality. Shame, he says, elicits contempt, derision, or avoidance on the part of other agents while guilt, in line with Wallace's analysis, elicits resentment, anger, or indignation. Further, though the claim is attenuated, Williams argues that “there is some truth in the idea that Homeric society was a shame culture.” Others have argued that shame-cultures elsewhere, in particular in early Confucian Chinese communities, also utilized a distinctive set of reactive attitudes and norms that govern

---

38 Ibid. 67
39 Ibid. 89
40 Ibid. 78
their appropriateness.\footnote{The bulk of this discussion is borrowed from Jane Geaney's “Guarding Moral Boundaries: Shame in Early Confucianism” in Philosophy East and West 54,2 2004: pp. 113-142}

A note of caution is in order here however. Shame is often thought in these cases to denote that the very same emotion is being picked out in varying historical and social contexts. I hesitate to follow through with the practice because it seems to presume a sameness where I have been at pains to argue that this sameness should not be presumed. If Psychological Constructivism is at least roughly a true account of what emotions are then it is unlikely that our contemporary use of the word shame picks out the same set of attitudes, eliciting conditions, thoughts and judgments, and behaviors corresponding with classically Greek or English translations of early Chinese Confucian uses of the term. I follow the authors in this discussion in using the word shame to describe these emotions with the understanding that this represents only a loose understanding of the emotions we are actually dealing with.

This concern over shame becomes ever more important when authors describe the differences between shame as typically understood and shame as it was understood in early Confucian Chinese communities. The two notions are different enough to force the question: why call these two different responses shame? Unless we presume that shame is a universal human emotion this nomenclature becomes arbitrary. In any case there is, it is argued, a stark contrast between shame as understood in western culture, including shame as understood in ancient Greece, and shame as understood in early Confucian societies.
Shame, as I noted above, is typically understood in the way that Bernard Williams has analyzed the emotion: shame is a negative self evaluation that stems from either the actual reactions of others (or from the 'right set of others') or from the hypothetical reactions of an internalized 'other' to our own behavior. The negative self evaluation is the judgment that one has failed to be as one is meant or expected to be. For example I can be justifiably ashamed when it is discovered that I behaved dishonestly (or when I imagine myself being discovered behaving dishonestly). Shame, importantly, is most often spoken about in visual terms and especially in sexual terms. We are ashamed to be seen naked, especially with our genitals uncovered. Gabriel Taylor argues that this is because shame involves having our true selves seen without our consent and our sexuality in part represents this true self; for Williams shame is visual because it implies a certain kind of loss of power to be exposed in this way by others. In either case shame is an evaluation of whole agents (and not particular acts) we should be ashamed of ourselves and the proper response to shame is to hide, to run away so as to avoid being seen.

Guilt, on the other hand, is something we feel about ourselves as a result of judging ourselves as having caused a bad state of affairs. We feel guilty when we

---

43 Shame covers more than merely acting appropriately as one could be ashamed, for example, of one's family or friends or even of being poor despite the fact that one could have behaved exactly as expected. We could be ashamed of our circumstances despite having done what is expected of us.
44 The matter for Williams regarding the nature of shame is also complicated by the fact that he thinks that some, but not all, of the work we now assign to guilt (compensating others for our wrongdoing) was carried out by shame for the Greeks. To the degree that he is right about this we should again be careful to consider our specifically contextual use of shame here to describe the Greek emotion in question. For Taylor see *Pride, Shame, and Guilt*. 
judge that we have done wrong. It is this sense in which guilt, but not shame, is morally focused. We can be ashamed at having behaved imprudently or rudely or immorally but we can only feel appropriately guilty at having behaved wrongly. Guilt is also less dependent on others than shame in the sense that we judge ourselves to have done wrong and doing wrong here can be described without reference or exposure to the reactions or possible reactions of others. Guilt is also independent of other non-moral norms. Slavery could be the norm in a community and yet it seems appropriate, to us, for a person owning slaves to feel guilty about it. Shame, at least on the analysis given by Williams and Taylor, does not appear to function independently from society in this way.

But shame in early China, according to the Jane Geaney, shares almost none of the features associated with guilt and shares only a few of the features associated with Classical Greek shame. Shame in early Confucian Chinese context, we are told, is not primarily referred to with visual metaphors nor does it have anything to do with exposure to a real or hypothetically internalized public. According to Geaney, shame in this context is best understood in terms of the blurring of boundaries. The objects of shame are agents or objects that represent blurred boundaries and the proper response to shame here is not to avoid being seen but instead to modify one's behavior in order to reassert the boundaries that have become blurred. Material goods, for example, are the most common proper object of shame because they are used to mark off status boundaries. When material goods are either enjoyed too much or not enough then
status boundaries become blurred. Commoner may look like an elite or an elite may look like a commoner. Corpses are also the object of shame, on this view, because they blur boundaries between living and dead and soul and body. Families must therefore treat corpses in ways that determine these blurred boundaries (stuffing the mouth with rice and the ears with jade for example reasserts the living purpose of these orifices and removes the shame). On my terms this “boundary-shame” bears more resemblances to our notions of disgust than to shame. For example there are concerns boundary shame shares with pollution that pertain more to the magical thinking of disgust than to shame as described above with Williams and Taylor. Boundary shame does seem to involve some degree of community perception as with Classical Greek shame. The motivation to call this emotion shame, in the face of such differences, is weak.

If this analysis of boundary-shame is correct however then we can here propose a different way of understanding the emotions involved in norm enforcement with the early Chinese case. Although the sense of shame here is quite different than the sense of shame invoked by Williams and Taylor they both seem to lack the distinct sense of moral responsibility that comes with an emotion like guilt and its pairing with resentment that Williams finds lacking in the Classical Greek conception of shame. The thought here being that a conception of guilt is paired with a conception of resentment and moral wrongness in cultures with these concepts. Insofar as this marks

---

45 The situation with the Greeks is of course complicated by the added fact that our own culture develops in part from theirs. Williams notes that this makes it more difficult to separate the historical development of our own moral concepts, which includes Greek ideas, from the Greek moral concepts themselves.
an important contrast for us, morally speaking, then we can see these two examples of
supposed shame-cultures as case studies that supplement the appeal to cross-cultural
reactive attitude data to support the constitutive claim for the reactive attitude
theorists.

Williams suggests in the case of the Classical Greeks and Geaney suggests for
early Chinese Confucian communities that cultures who use shame as the primary
reactive attitude to enforce social norms lack a distinct notion of moral responsibility.
This is because they do not make use of guilt to enforce their norms. The guilt-
resentment attitudes are important because only the guilt-resentment attitudes are
paired with a distinctly moralized notion of wrongness. Feeling guilty requires
thinking that we have behaved wrongly. ETRA seems poised to deliver the best
explanation for the differences between the evaluative concepts of the classical Greeks
and early Confucian Chinese communities and our own moral concepts. We do
however have several options if we want to try to explain these differences: (a) shame-
cultures are more primitive than guilt-cultures (they will or would eventually develop
into guilt-cultures); (b) shame serves an epistemological and motivational purpose in
shame-cultures but an incomplete one. Shame is a poor proxy for resentment,
indignation, or guilt. It is not until guilt is fully utilized that a culture gets access to the
true evaluative concepts pertaining to moral responsibility and in shame-centric
cultures moral accountability is either ignored or reinterpreted under different
regulative norms. Finally we can appeal to ETRA and say that (c) shame, in its
different senses, can constitute different sets of evaluative concepts pertaining to agency that are used to enforce social norms of cooperation. Although other cultures, like our own, use guilt and moral responsibility to enforce their norms, shame-cultures appeal to distinctively different notions of agents and what might loosely be called attributability, as a result of their understandings of the appropriateness conditions for shame.

If Williams is right then the Classical Greeks should be understood as a culture in which a cluster of reactive attitudes including an emotion resembling shame were primarily responsible for enforcing norms and ensuring social cooperation. If Geaney is right then early Confucian Chinese communities relied on a different emotion, somewhat resembling shame, and on the norms that govern it, that are very different from resentment, indignation, and guilt as typically understood. If this analysis is right then we have a successful test case for ETRA. The worry here is not merely that different cultures will differ on whom is responsible or why since this implies that a concept of responsibility is shared in these cultures. The claim is that differing cultures will not only have differing practices of norm enforcement, utilizing different reactive attitudes, but that what makes those attitudes appropriate will also diverge from our standard notions of the core concepts of responsibility and responsible agency.

If Williams is wrong about the Classical Greeks, if they are best seen not as a 'shame' culture but as more in line with having our set of reactive attitudes (albeit in a different cultural context) then it is up to the reactive attitude theorist to find another
candidate for ETRA. If none are found or if, even in those cultures, clusters of concepts similar to our own can be found then the constitutive thesis is cast into serious doubt. In any case it appears a fruitful avenue for the reactive attitude theorist to draw intuitive support for the constitutive claim.

**Conclusion**

Sentimentalism, as I define it, requires having a theory of (at least some) evaluative concepts in which those evaluative concepts are constituted at least in part by—the content of the evaluative content is at least in part irreducibly determined by—emotional responses. The last two chapters have been a hunt for arguments that support this core claim. In the last chapter I canvassed general second-order Sentimentalist theories and came out wanting. In this chapter I applied a similar strategy to proponents of reactive attitude theories of emotion. This family of theories does not, on its face, aspire to vindicate Sentimentalism about all or most evaluative concepts. Its sights are narrower and constrained to showing only that our concepts of responsibility and responsible agency are intimately linked, constitutively so, with some set of special reactive attitudes.

In canvassing reactive attitude arguments for constitutiveness I found some familiar problems. Strawson, for example, makes the risky assumption that the reactive attitudes are universally shared because they are basic. Furthermore he goes on to assume that universality implies constitutiveness. Both of these claims are doubtful. Wallace does better by dropping the idea that the reactive attitudes are
universally shared. However, his analysis of 'deep' responsibility, the kind of responsibility that sees reactive attitudes as constitutive, fell short. Although Wallace claims that we fail to fully understand the reasons why we believe that agents act wrongly, saying that an agent has acted unfairly and that the agent has reflective self control says much about the capacity of the agent and the failing for which she is being held responsible. The expressive content of the reactive attitudes may be left out of this picture but to assume that leaving this out fails to capture the content of our concepts of responsibility and responsible agency is to assume the constitutive claim not argue for it. Wallace does propose an interesting anthropological test for our reactive attitudes and it is here that I think the arguments have most promises.

The allure of psychopaths, I also concluded, should be directed to the epistemological role for emotions and to questions about whether or not psychopaths themselves are morally responsible agents. Psychopaths make for interesting test cases for the constitutive claim only in combination with Wallace's use of shame-cultures into what I called the Empirical Test for Reactive Attitudes (ETRA). This was because psychopathic agents are strongly motivated, but apparently fail, to acquire the moral concepts of their respective cultures. This implied the epistemological role on its own. However, if psychopaths motivate the epistemological role then shame-cultures pushed us farther still. If agents in shame-cultures are being led to form different evaluative concepts as a result of their appeal to the conditions of appropriateness for shame then we are left with a small mystery if shame were only playing the
epistemological role. We would need an explanation for why shame should provide access to conceptions of agency and norms of social cooperation that were different from our own. It seems imperialistic to claim that these conceptions are simply mistaken or primitive. ETRA provides the better response, I argue, by appealing to the constitutive role as the best explanation. However, even in the best case scenario I think ETRA provides us with some motivation to continue pursuing the constitutive claim. Though this is far from showing that the constitutive claim about reactive attitudes and responsibility is true, it lends plausibility where plausibility is needed. If facts about moral responsibility were response-independent then we should expect more uniformity about the nature of the concept. Uniformity, however, seems to vary depending on the degree to which an agent has access to resentment, indignation, and guilt. This appears true not only with the clearly aberrant psychopath but also in otherwise normal humans in shame-cultures.

Sensible Sentimentalism needs to be able to accommodate several features. It would be best if Sentimentalists do not commit themselves to contentious empirical claims. A commitment to basic emotions is just such a claim. While the existence of a set of universally shared and independently identifiable emotions would be useful to Sentimentalists the prospects for such a view do not look good. Furthermore this view of basic emotions is harder to square with the Psychological Constructivist theory of emotion I have been suggesting best accounts for our emotional experiences. On that view emotions are not cognitively impenetrable evolutionary affective modules as
they were on views like those of Paul Ekman. While some components of emotional experience, like core affect, do appear universally shared and phylogenetically primitive these components fail to capture, by themselves, the richness of our actual emotional lives. Sensible Sentimentalism would be better served, I argued, by incorporating a view like Psychological Constructivism in order to explain emotions. While this limits the potential scope of moral claims it need not undermine Sentimentalism generally. Evaluative discourse can be explained by appeal to emotions of this sort, as I suggested at the end of the last chapter.

I ended this chapter by showing how Sentimentalists could begin to provide an argument in support of the constitutive role for emotions within a theory of moral responsibility. Although this argument is by no means bullet proof it does go farther than traditional Sentimentalist arguments in terms of motivating the thought that emotions could be constituting at least some of our evaluative concepts. Sentimentalism is on its way to sensibility.
References


D'Arms, Justin. Stanford Encyclopedia entry on *Envy* available at: (http://plato.stanford.edu/entries/envy/)


James, William. “What is an emotion?” Mind 9, 1884: pp. 188-205.


-“Affective consciousness: Core emotional feeling in animals and humans” *Consciousness and Cognition* 14 2005: pp. 30-80.


