Cognition and Emotion: A New Approach

by

Paul Andrew Jamieson

A thesis submitted in conformity with the requirements

for the degree of Ph.D.

Graduate Department of Philosophy

University of Toronto

Cognition and Emotion: A New Approach

Ph.D., 2001

Paul Jamieson, Department of Philosophy, University of Toronto

Emotions are more cognitively complex than philosophers have typically thought. That is the simple claim I argue for in this thesis. And while it is a simple point it has profound implications. Most importantly, it means that philosophy must expand its methodology beyond conceptual analysis—its favourite method of studying the emotions—and align itself with the flourishing empirical study of emotion.

In the first chapter I offer a selective history of the philosophy of emotion intended to show how philosophy's current understanding of emotion has developed. The main character here is Aristotle. His complex understanding of the emotions prefigures much of the typical modern account. He saw, for example, how emotions are intimately tied to particular classes of belief and developed a detailed formal analysis of these ties.

For all the power of Aristotle's analysis, however, it does have its problems. In the second chapter I identify a modern, distorted version of Aristotle's account that I call 'hyper-cognitivism.' The essence of hyper-cognitivism is a myopic focus on emotion's cognitive elements that is, strangely, coupled with an ignorance of the complexity and variability of those elements. This ignorance, I argue, stems largely from a misguided methodological reliance on conceptual analysis that has led many philosophers to ignore what the empirical sciences tell us about emotion.

This is a particularly unhappy oversight because the empirical study of emotion is currently undergoing a long overdue boom. Most excitingly, the neurosciences have begun to

study the neural foundations of emotion. In the third chapter I offer a brief sketch of the picture of emotion that is emerging as a result. In particular, I focus on evidence that suggests that many of our emotions are subserved in the brain by discrete 'emotion systems.' These systems possess a range of basic cognitive capacities that are capable of operating independently of higher brain systems that mediate more complex forms of cognition. In the fourth chapter I argue that philosophy should incorporate these insights into a unique form of description that captures the contribution of these basic forms of cognition to the production of emotion.

Acknowledgments

It is a pleasure to thank the many people who have supported me during the writing of this thesis. I am particularly indebted to my supervisor, Ronnie de Sousa. Ronnie was unfailingly generous with his encouragement, and his many helpful criticisms were consistently delivered with a flair and humor that enlivened every meeting. All should be so lucky to have a supervisor as dedicated and professional.

I have also benefited greatly over the years from the fellowship of my colleagues in the philosophy department at the University of Toronto. I am especially grateful to Patricia Arnold and Jennifer Gibson. Their steady friendship eased many difficult times.

Outside of philosophy my family and friends provided all the loving support and distraction necessary to the completion of a long and absorbing project. Lauren Greene-Roesel, in particular, has made the last two years infinitely easier for me in more ways than she can know. Finally, I owe the greatest thanks to my parents, John and Sheilah Jamieson, for their unwavering love and support. This thesis is as much theirs as mine.

Table of Contents

List of Appendices

Chapter One: A Brief History

Introduction

A cursory review of mainstream modern philosophy of mind reveals a singular fact: serious treatments of the emotions are conspicuously rare. They seem, to borrow Sue Campbell's apt phrase, to have "fallen off the map of the mind."[1] Fortunately, the same cannot be said of the history of philosophy. Philosophers of the past have had a great deal to say about the human passions. They have talked at length about their causes, their nature, and their role in the production of behaviour. They have considered the relation of emotion to thought, rationality, and the ethical life. They have offered methods to cure us of our destructive passions, and to inculcate in us those most beneficial to our well-being.

It follows then that any history of the philosophy of emotion must be highly selective if it is to remain a reasonable length. The history offered here thus focuses almost exclusively on what only a few philosophers, and one biologist, have said about the relationship between emotion and cognition. Of course, philosophers of the past seldom used the leaden term "cognition" when considering the 'rational' side of emotions. They instead typically employed more familiar and euphonious terms like "thought" and "judgement." In my history, however, I will use "cognition" rather freely, for two reasons. First, it is a handy catch phrase, allowing an expositor of historical theories to smooth over unimportant differences and connect ancient observations with

data from the modern laboratories of experimental psychology and cognitive science. Secondly, I hope that by repeating the phrase enough it will grate on the reader's sensibilities to the extent that they see that "cognition" is *nothing more* than a handy catch phrase, and are driven to demand more precise definitions of the processes it confounds. But more about this below.

A final note. Aside from their historical influence, my motivation for discussing the theorists that I do is my belief that for the most part they got things right. Plato is perhaps the exception here. His assignment of emotion to a distinct faculty of the mind is flat wrong, but the idea has been so influential it must be noted. Happily, though, it turns out that Plato actually held a somewhat subtler view of emotions; the 'distinct faculty claim' is found largely in his metaphorical descriptions of the human subject, not in his actual explanation of how emotions are structured. As for the others—Aristotle, Descartes, and Darwin—in addition to being highly influential, they were often astoundingly prescient. Aristotle's account of the cognitive structure of emotions contains nearly all the features of the typical modern analysis; Descartes conception of the function of emotions prefigures a prominent modern view that emotions work to promote a beneficial alignment of body and mind in response to highly significant stereotypical situations; and Darwin's studies of emotional expressions have served as the basis for the modern evolutionary investigation of emotion, a field that has in turn confirmed many of his original observations.

---

[1] Campbell 1997, p. 13.

Plato: Separation, Opposition, and the Truth of Feeling

I turn first to Plato not as a theorist but as a source for two themes that have strongly influenced Western thinking about the emotions: the *separation* and *opposition* of reason and passion. In both common and philosophical myth the poles of reason and passion have long been taken to mark the boundaries of the human economy of mind. When Plato divided the soul into its three proper parts—the rational, appetitive, and emotional—he quite literally located reason at the top of the body.[2] Emotion and appetite were encased in the breast, with emotion nearer to the head than appetite so that it might more easily "join with [reason] in controlling and restraining the desires." [3] This figure and its attendant metaphors have shown admirable endurance and influence. The conception of the human mind as divided into distinct faculties of reason and passion is unarguably commonplace. As for the dynamic of these opposing features of mind, the dominant trope is again due to Plato. Passion is "a contentious and combative element which frequently causes shipwreck by its headstrong violence."[4] Passion thus stands to reason as a wild horse stands to its charioteer: a necessary motive force, but one that is dangerous if left unchecked by a more ordered and cautious power.[5] It would be wrong, however, to read Plato's rhetoric as accurately embodying his theory of emotion and its relation to reason. Even a brief examination of his dialogues reveals a more subtle account of the natures and relations of passion and reason. As an example, I focus here

---

[2] An argument for the tripartite nature of the soul is found in *Republic*, 435ff. Plato locates reason "at the top of the body" in *Timaeus*, 89e.
[3] *Timaeus*, 69d.
[4] *Laws*, 9.863b.
[5] The imagery of the horses and charioteer is found at *Phaedrus*, 246ff.

on only a single section within the *Philebus*, a dialogue concerned with determining which state of the soul—pleasure or thought—will render one's life most happy.

In the midst of this dialogue Socrates puts the question to his interlocutor Protarchus: "Shall we say that...pains and pleasures are true or false? Or that some are true, and others not?"[6] The investigation that follows works carefully through the nature and relations of opinion and feeling. It is first noted that pleasures and pains can be construed as 'true' in the sense that it is not possible to think one is experiencing a particular pleasure and yet not really feel it; the pleasures of dreams are as 'true' as the pleasures of waking life. Given the 'truth' of feeling in this sense is it possible for feeling to also be false? If not, the 'truth' of feeling is not truth in even the weakest sense of common usage for there is no contrast with falsity. The question for Socrates thus becomes: how might feeling be false?

On the grounds that *opinion* is the paradigmatic bearer of truth and falsity, Socrates now proceeds by comparing feeling and opinion. They are alike in the first place, he notes, in that just as there are the similar states of *holding* an opinion and *feeling* a pleasure, there is likewise always a thing *about which* an opinion is held and *in which* a pleasure is felt. Moreover, as is true of feeling, no one is ever wrong when they think they hold an opinion regardless of the truth or falsity of that opinion; there is no difference between my believing I am of the opinion x and actually being of the opinion x. Even more promisingly, Socrates observes that "we often experience pleasure in association with an opinion that is not right, but false."[7] The relation between opinion

---

[6] *Philebus*, 36c.
[7] *Philebus*, 37e.

and feeling is thus stronger than merely sharing similar characteristics; there is sometimes a *causal* relationship between the two. Noting these similarities and connections, Socrates now poses the essential question: "How is it that whereas we commonly find opinion both true and false, pleasure is true only, and that though in respect of reality holding an opinion and feeling a pleasure are on the same footing."[8]

By way of an answer Socrates turns to the source of opinion's veracity. If feeling and opinion share the noted similarities and are sometimes causally connected then perhaps feeling can be false in the way that opinion can be false. So how can opinion be false? Socrates argues that the source of an opinion's potential falsity lies in the possibility that "it might sometimes have reference to what was not a fact, either of the present, the past, or the future."[9] We thus now have two potential ways in which feeling might be false. First, a pleasure *caused* by a false opinion might be said to itself be false in virtue of this association. Socrates briefly considers this possibility but Protarchus offers the obvious objection that "we call the opinion false, but the pleasure itself nobody could ever term false."[10] Socrates then considers the second possibility that feeling and opinion are true and false in exactly the same way, and without much in the way of supporting argument, reaches just this conclusion: "...though anyone who feels pleasure at all, no matter how groundless it be, always really feels that pleasure; yet *sometimes it has no reference* to any present or past fact, while in many cases, perhaps in most, it has reference to what never will be a fact."[11] *Lack of reference to fact* is thus the source of

---

[8] *Philebus*, 37b.
[9] *Philebus*, 40d. Socrates leaves unanalyzed the notion of an opinion's "having reference to" a fact.
[10] *Philebus*, 38a.
[11] *Philebus*, 40d.

feeling's falsity, as it is the source of opinion's falsity. Moreover, Socrates claims, this principle holds not just for simple pleasures and pains, but is also true of our more complex emotions like fear and anger.[12]

Unfortunately, Socrates' argument in this section is truncated and weak. The conclusion is asserted without argument and the central notion of 'reference' is left opaque. The value of the *Philebus'* arguments concerning opinion and feeling, however, lies in the posing of the question: can emotions be true and false? By asking the question and answering positively—no matter the answer and its supporting arguments are not particularly satisfying—Plato belies the simplistic separation and characterisation of reason and passion embodied in the metaphors and tropes he created. Truth and falsity are judgements applied within the arena of rationality and to argue that feelings and emotions are capable of truth and falsity is to bring them under its purview.

We can thus take from Plato the germs of four important but underdeveloped themes:

1. Emotion and reason are separate and distinct faculties of the mind.
2. Emotion and reason can come into conflict when emotion motivates us to act against the better judgement of reason.
3. Despite their capacity to conflict in this way, emotion and reason share some defining characteristics. Most importantly, our emotions and our opinions both involve *reference to fact*. This opens up our emotions to judgements of truth and falsity.

---

[12] *Philebus*, 40e.

4. Our emotions and opinions are further related in that opinions can sometimes be the cause of emotions.

Aristotle: The Cognitive Structure of Emotion

Aristotle's theory of emotion is found in its most complete form in the *Rhetoric*, and the reason for its inclusion in a treatise on the art of persuasion is made apparent in Aristotle's general definition of the passions: "Passions...are all emotions whatsoever, on which pain and pleasure are consequent, by whose operation, undergoing a change, men differ in respect to their decisions."[13] The complete rhetorician must therefore have the capacity to manipulate his audience's passions, for to manipulate these is to potentially manipulate their judgements. Aristotle's approach is thus founded on pragmatic concerns and his analysis proceeds with an eye on the practical. Emotions are to be studied along three dimensions:

> ...it will be fitting to divide what I have to say, respecting each [emotion], into three considerations; to consider, respecting anger, for example, how those who are susceptible of anger are affected; with whom they usually are angry; and on what occasions. For granted that we be in possession of one, or even two of these points, and not of them all, it will be impossible for us to kindle anger in the breast.[14]

As we proceed it is helpful to keep in mind this pragmatic bent to Aristotle's subsequent analysis. Given that his concern is to allow the orator to induce particular emotions in their audience—to "kindle anger in the breast"—Aristotle's account must penetrate to the

---

[13] *Rhetoric*, II.ii.vii.

[14] *Rhetoric*, II.ii.viii. In *The Nichomachean Ethics* Aristotle identifies feelings (*pathos*) as one of "the three modifications that are found in the soul," the other two being faculties (*dunamis*) and dispositions (*hexis*):

ways in which our emotions *actually work.* He must, for example, be more concerned with the psychological mechanics of the passions than with conceptual analysis, i.e., with detailing how emotion words are commonly used.[15] Why, however, be concerned with supplying this information to a rhetorician—as opposed to a physician—who can act upon us only through argument? As noted above, the motivation here is partially Aristotle's belief that emotions affect our decisions—the ultimate end of rhetoric. The more interesting implicit claim, though, is that just as our "decisions"—i.e., our considered judgements and evaluations—are susceptible to the influence of our emotions, our emotions are themselves susceptible to the influence of argument. An important upshot of this claim is that emotions cannot simply be bodily sensations. No argument, however persuasive, can alleviate a toothache's sting or the ache of a wound. Emotions must rather be *of the nature* of those things that are paradigmatically susceptible to the influence of argument, namely beliefs, judgements, and evaluations. For Aristotle, there thus exists a chain of effect stretching from argument to emotion to judgement; emotions both act upon and are acted upon by reason. How this can be so becomes clear in the details of Aristotle's analysis of our individual emotions.

The tripartite structure of Aristotle's suggested analysis parallels his conception of the constitution of the particular emotions. Roughly, Aristotle holds that emotions involve three distinct aspects: (1) a *cognitive* element, some particular state of mind, belief, judgement, or evaluation;[16] (2) a *conative* element, usually a desire to perform

---

"By feelings I mean desire, anger, fear...and in general all conditions that are attended by pleasure and pain" (*Nic. Eth.* II.v.).

[15] Nussbaum 1994, p. 82.

[16] What is the precise nature of emotion's cognitive element? Martha Nussbaum (1994, p. 84) notes that throughout his definitions in the *Rhetoric* Aristotle shifts in a loose fashion between weaker 'appearance'

some particular action where the nature of that action is partially definitive of the emotion; and (3) a *hedonic* element, i.e., a pain, pleasure, or some particular mix of the two. Anger, for example, is defined as "a desire accompanied by pain of a revenge which presents itself, on account of an apparent slight from persons acting toward one's self, or some of one's friends, unbecomingly."[17] Fear is "a sort of pain or agitation, arising out of an idea that an evil, capable either of destroying or giving pain, is impending on us."[18] Pity is "a sort of pain occasioned by an evil capable of hurting or destroying, appearing to befall one who does not deserve it, which one may himself expect to endure, or that some one connected with him will; and this when it appears near."[19]

The first point of note in these definitions is the looseness and variation shown in detailing which elements of a particular emotion are *constitutive of* that emotion and which are merely necessary—or necessary and sufficient—conditions for its appearance. This distinction is typically thought an important one and much philosophical effort has been spent attempting to sort out the relative roles of emotion's components, especially in relation to its cognitive elements. In modern debates the question is generally posed in terms of a causal/conceptual distinction. Are certain cognitions the *separate* causes of subsequent and distinct emotional states, or are they partially constitutive of them in the

---

words like *phantasma*—the appearance of the sun as a foot wide is a *phantasma*—and stronger belief words like *dokein* and *oiesthai*. For an appearance to graduate to the cognitively stronger state of belief an act of assent is required; the subject must adopt some appropriately positive attitude towards the appearance. So where does this leave us? Even a casual reading of the definitions shows clearly that for Aristotle the cognitive elements of our emotions are *typically* more complex than simple images. He does, however, allow that emotions are sometimes caused by cognitively simpler states like *phantasmata*. I will say more about this below. For the distinctions Aristotle draws between *phantasmata* and the more complex states of belief, knowledge, and conception, see *On The Soul*, 3.3.

[17] *Rhetoric*, II.ii.i.

[18] *Rhetoric*, II.v.i. Cf. Plato's brief definition of fear in the *Laws*: "...the special name for anticipation of pain being *fear*" (I.644d).

[19] *Rhetoric*, II.viii.i.

sense that the concept of a particular emotion like fear *logically presupposes* the occasioning of a particular cognition, say, the judgement that I have been slighted.[20] The two views seem irreconcilable in light of Hume's principle that causal relations can hold only between items that are logically distinct. So how are we to interpret Aristotle on this point? At the very least it is clear that while each distinct emotion is held to involve some combination of passion's three basic elements, Aristotle nowhere in the *Rhetoric* explicitly argues for an overarching formal account that places each element in a fixed relation and ratio to the others. I will suggest below, however, that we should read Aristotle as falling on the conceptual side of the modern debate. The various cognitions, desires, and feelings involved in emotions are not, for Aristotle, the contingent prerequisites of further and distinct emotional states, but rather are constituent of those states. At this point, though, I want to turn to the question of how emotions are individuated. The general outline of Aristotle's theory of emotions comes out most clearly in tracing out what affords each emotion its unique identity. I turn first to Aristotle's account of anger.

Anger is for Aristotle partly individuated by the *formal structure* of its cognitive component—the cognition 'that one has been slighted.' I use the term "formal structure" here to emphasise that in Aristotle's definition 'slight' is used as a *genus* term. Anger's defining causal cognition—the judgement that 'I have been slighted'—is only a formal placeholder standing in need of instantiation with particular concrete cognitions of the proper type. That is, *actual* instances of anger are occasioned by specific cognition *tokens* of the *type* 'I have been slighted.' What more specific form will these tokens take?

---

[20] For modern arguments on the conceptual side see Bedford (1957) and Thalberg (1977).

Aristotle notes that there are "three species of slight: contempt, vexatiousness, and contumely."[21] Thus people can "feel anger towards those who laugh at them excessively, and gibe, and scoff at them, for these treat them with contumely; with such also as hurt them in all particulars, of such a nature as are tokens of contumely."[22] A cognition that I have been the butt of excessive laughter is a token of the type 'I have been slighted' and thus drives me to anger. At the formal level then, anger is for Aristotle a narrowly constrained concept in that its eliciting cognition can be characterised by a very specific genus or type term.[23] Of course, this narrow construal still allows for a wide range of concrete instances of anger, each differing from the others in virtue of what would now popularly be called the 'content' of the tokened cognition. What ties all the various instances together as instances of anger, however, is that in virtue of their specific content these cognitions are capable of being subsumed under the type description 'I have been slighted.'

The defining cognitions of our emotions are also formally structured in the sense that they can be analyzed into unique and discrete 'dimensions.' Put simply, the cognitions that are partially constitutive of our emotions have their own 'parts,' these being specific dimensions along which the cognition can vary. Fear and pity, for example, are identical in hedonic aspect, are occasioned by evils of precisely the same nature, and differ only in the subject's belief about whom the evil will affect. Fear arises when the evil threatens one's self, while pity arises when it threatens an undeserving

---

[21] *Rhetoric*, II.ii.iii.
[22] *Rhetoric* II.ii.xiii.
[23] This idea finds modern expression in the notion that emotions are individuated by their *formal objects*, a claim I will discuss in the second chapter.

other. Thus Aristotle notes that "to speak generally, all those things are to be feared, which, happening or being likely to happen in the case of others, excite compassion."[24] The cognition that causes both fear and pity might then be represented as

S will be visited by a destructive evil

where S marks what could called a 'Subject' dimension. It is, simply, a placeholder that can take one of only two values: 'I' or 'Other.' In the case of fear and pity the identity of the instantiating value determines the identity of the attendant emotion. Note, moreover, that we might more completely formalise the cognition that Aristotle holds as definitive of fear and pity as

S will be visited by a Q of type T.

Here, Q marks what might be called a 'Quality' dimension. Again, it is simply a placeholder capable of taking on one of two values: 'Good' or 'Evil.' 'T' marks a dimension of variation which qualifies the impending good/evil denoted by Q. "People do not fear every evil," Aristotle notes, "but people fear all those evils whose effect is either a considerable degree of pain, or destruction."[25] The value of T will thus fall on a continuum lying between the poles of malignancy and benignness, and fear will only arise when the cognition affords the impending evil the necessary degree. Such a formalism is, of course, alien to the style of Aristotle's theory, but not to its spirit, for it brings to the fore a central feature of his analysis of the emotions. Emotions are for Aristotle highly structured states in that they are constructed from distinct components—

---

[24] *Rhetoric*, II.v.xii.
[25] *Rhetoric*, II.v.i

cognitions, desires, and feelings—that are themselves structured in the sense just developed.

This aspect of Aristotle's theory is emphasised further when we look at the other differentiae of our emotions: the conative and hedonic elements. In anger, the desire for revenge is part of what uniquely identifies the emotion.[26] Our other emotions also contain defining desires.[27] The difference, for example, between the related and similar emotions of anger and hatred is marked in part by the 'formal' ends of their partially constitutive desires: "[anger] is a desire of inflicting pain on its object, [hatred] of doing him deadly harm; for the angry man wishes to be felt, to him who bears hatred this matters not."[28] Again, and more obviously here, anger's definitive desire—the desire for revenge—is to be read as a formal requirement. The precise content of actual tokened desires is allowed by Aristotle to shift as circumstance dictates. I might wish any of a number of evils upon the object of my anger but they will, according to Aristotle, be (1) of a degree less than deadly, and (2) accompanied by a desire that their victim recognise me as their author.

This desire for revenge, while necessary to anger's identity, also interestingly complicates its third differentia—a painful hedonic aspect. Aristotle sets anger off from

---

[26] Hence Aristotle is careful to note the difference between easily confused goals of revenge and punishment: "punishment is for the sake of the sufferer, but revenge for that of the person inflicting it, in order that he may be satiated" (*Rhet.* I.x.xvii). So while similar in effect, the desires for revenge and punishment differ importantly, and anger can only move us, at least directly, to vengeance.

[27] The conative aspect of emotion generally seems to be of less importance in defining an emotion's identity than the cognitive. While all emotions discussed by Aristotle involve particular desires to varying degrees, only his definitions of anger, hatred, and friendliness actually mention a constitutive desire. The definitions of fear, shame, pity, indignation, envy and gratefulness all lack specific mentions of desire, though subsequent discussion of each touches on 'attendant' desires.

[28] *Rhetoric*, II.iv.xxx.

hatred in part by the presence of pain: "...anger is attended by pain, hatred is not."[29] While anger is by definition a pain, however, the prospect of revenge mixes with that pain a pleasure derived from hope: "...there is a sort of pleasure consequent on all anger, arising out of the hope of avenging oneself."[30] It follows, moreover, from the necessity of anger's desiring revenge that "there is no one who feels anger where the object seems impracticable to his revenge; nor with those far their superiors in power do men feel anger at all, or if they do, it is in a less degree"[31] Just as the possibility of revenge colours anger's pain with pleasure, the impossibility or unlikeliness of revenge moderates our anger *in toto*.

This last point is an example of what for Aristotle is a central feature of emotion: the conceptually distinct components of our emotions are not separable in practice. Our ocurrent passions ebb and flow *as a whole* in measure with the course of their individual, constitutive cognitions, desires, pains, and pleasures. Discussing 'placability,' the opposite of anger, Aristotle notes that "men are placable, when in a frame of mind contrary to the feeling of anger; thus in amusements, in mirth, in festivity, amid rejoicings...in a word, when in a state of freedom from pain."[32] In short, when we are in states disposed toward the absence of pain we are accordingly disposed to anger's opposite, placability. Similarly, just as the impossibility of revenge diminishes our anger, satisfaction of anger's defining desire also moderates the emotion as a whole: "Men are thus disposed [to placability] if they have convicted the object [of their anger], and if he

---

[29] *Rhetoric*, II.iv.xxx. No reason is provided for why hatred lacks a painful aspect. Aristotle generally seems little concerned with close analysis of the nature of emotion's hedonic element, and is usually content to merely note the presence or absence of pains and pleasures.
[30] *Rhetoric*, II.ii.iii
[31] *Rhetoric*, I.xi.ix.

suffered a greater ill than they, with all their anger, would have themselves inflicted; for they think they have gotten...their revenge."[33] Finally, anger subsides in the absence of an essential element of its cognition—the judgement that I have suffered an *undeserved* slight. Thus anger is not felt by the subject "if they are aware that they are themselves unjust, and suffer deservingly; because anger is not felt at what is just."[34] Our other passions are similarly dependent upon the co-existence of emotion's three basic elements. The lesson here, I suggest, is that we should read Aristotle as conceiving of an emotion as a complex unified state, the integrity and identity of which depends upon the presence and nature of its three constitutive elements.

I would further suggest that this understanding places Aristotle firmly on the 'conceptual' side of the modern debate about the proper role of cognition in emotion, though with an important amendment. The cognitions Aristotle mentions in the definitions of our various emotions are not prior and distinct states that cause subsequent and distinct emotional states. They rather *are* those emotions, when properly combined with the right desires and feelings. I would suggest, however, that Aristotle likely conceived of the notion 'constitutive of' rather differently than modern theorists of emotion. He nowhere speaks, as many moderns do, of the presence of a particular emotion *logically entailing* the occurrence of a particular cognition. This is what is most usually meant when it is claimed that particular cognitions are *constitutive of* particular emotions.[35] I suggest that this difference stems in part from the decidedly pragmatic bent

---

[32] *Rhetoric*, II.iii.xiii.
[33] *Rhetoric*, II.iii.xiv.
[34] *Rhetoric*, II.iii.xv.
[35] E.g., see Gordon 1987.

of Aristotle's analysis. As noted above, Aristotle was concerned with detailing the actual workings of our emotions, as opposed to merely analysing common conceptions of emotion. Logical entailment, however, is primarily a feature of concepts and thus does not necessarily figure in an analysis of what emotions really are, for there is no immediately apparent reason for believing that the constitution of our emotions could not have been different than they are.[36] Analysing emotions as concepts, though, creates a space for the introduction of logical entailment, and I would argue that our ability to identify instances of logical entailment at *this* level of analysis tells us more about our beliefs about emotion than about our actual emotions. As I will say more about this in subsequent chapters, I want to turn now to the penultimate feature of Aristotle's analysis with which I will deal in this section: his conception of emotions as *intentional states*.

As stated in his discussion of the proper analysis of the emotions, Aristotle holds that a complete account of any emotion must include a discussion of its proper and usual objects. In his various analyses our emotions are always directed *toward* some object. I am not simply "angry" but "angry at," just as I always feel "pity for" and am "jealous of." Consider one of the defining differences between anger and hatred. Aristotle notes that "it must be that he who is affected by anger, is so affected invariably *towards* some

---

[36]I draw this point from an observation of Hume's: "According as we are possess'd with love or hatred, the correspondent desire of the happiness or misery of the person, who is the object of these passions, arises in the mind, and varies with each variation of these opposite passions. *This order of things, abstractedly consider'd, is not necessary. Love and hatred might have been unattended with any such desires, or their particular connexion might have been entirely revers'd.* If nature had so pleas'd, love might have had the same effect as hatred, and hatred as love. I see no contradiction in supposing a desire of producing misery annex'd to love, and of happiness to hatred. If the sensation of the passion and desire be opposite, nature cou'd have alter'd the sensation without altering the tendency of the desire, and by that means made them compatible with each other" (*A Treatise of Human Nature*, II.ii.vi; my italics).

individual...but not towards mankind."[37] Hatred, on the other hand, "may be borne even to whole classes; for everyone hates the character of a thief and an informer."[38] The difference in the potential targets of anger and hatred stems from the nature of their respective constitutive cognitions. Anger arises from the perception of having suffered a slight, and given that mankind is not an entity capable of slighting me I cannot feel anger towards it. Hatred, however, involves our conceiving a person "to be of a certain description."[39] As I can conceive of mankind, or lesser classes, under a certain description—say as "indifferent to my suffering"—it follows that I can hate mankind as a result. The potential range of an emotion's objects is thus partially tied to the content of its defining *formal* cognition in that an emotion's object must be capable of bearing that description under which it is conceived of by the subject.[40] Thus we cannot pity those who we conceive of as being immune to an impending evil, nor can we be angry with those who slight us with justification. The potential range of an emotion's object is similarly limited to those which are capable of satisfying that emotion's definitive desire; thus our desire in anger that the victim of our revenge recognise us as author of their misfortune disallows anger toward the insensible and the dead.[41]

---

[37] *Rhetoric*, II.ii.ii.
[38] *Rhetoric*, II.iv.xxx.
[39] *Rhetoric*, II.iv.xxx.
[40] The notion 'capable of bearing a description' bears discussion. In one sense, numerous objects are logically or conceptually incapable of bearing a given description: numbers have no flavour just as ideas have no weight. Clearly, though, there is no limit to the descriptions that objects are capable of being *conceived of* as bearing. So it is not that certain emotions *can't* be felt towards certain objects, but rather that they *shouldn't* be felt. Thus Aristotle quotes Virgil's judgment of Achille's anger at the dead Hector: "In his madness he is vexing a senseless clod" (II.iii.xvi). Achilles is clearly *capable* of hating Hector even though Hector is dead, but his anger is irrational because the object of his anger is incapable of satisfying one aspect of his anger's defining desire. For a modern account of the relations between emotions and their objects that is very much in this Aristotelian vein, see de Sousa 1987, pp. 108-139.
[41] *Rhetoric*, II.iii.xvi. Just as anger towards superiors is always mitigated by the impossibility of revenge.

Clearly, therefore, Aristotle conceives of emotions as being *intentional states* in the common sense of their always being *directed upon* or *aimed at* a particular object. Given this characterisation, a central question now arises. How does Aristotle explain the object-directed, intentional nature of our emotions? While a full answer here would take us too far afield, I want to look briefly at one possible account, as doing so will bring to the fore an undeveloped but important feature of Aristotle's account.

I draw here upon Victor Caston's account of Aristotle's theory of intentionality.[42] Very briefly, Caston argues that Aristotle grounds his account of the intentionality of *all* mental states in a straightforwardly causal account of perception and sensation. At the most basic level, when a subject perceives some object or state of affairs they undergo a corresponding bodily change in the "central organ" that is directly caused by the object of perception. What makes this bodily change a perceptual representation *of* that object is that the *proportions* of that object are *preserved* in the proportions of the bodily change.[43] Aristotle is suggesting here a mental form of analogue representation similar to that which allows compact discs to 'represent' the music they encode by preserving the original music's auditory and temporal magnitudes in the pits and lands that are read by a laser. Of course, an account of this form, however developed, cannot itself explain intentionality, since it implies that the 'representations' produced in this way are infallible; analogue representations are *always* about whatever caused them.[44] One of the

---

[42] Caston 1998.

[43] "How, when a person thinks of larger things, will the fact that he thinks of them differ from [his thinking of] smaller things? For everything inside is smaller, just as the things outside are proportional also. Perhaps just as something distinct in him can be taken to be proportional to the forms, so to [there will be something proportional] to the intervals" (*On the Soul*, II. 452b11-16. In Caston 1998, p. 262.)

[44] "...for perception of the special objects of sense is always free from error" (*On The Soul*, 3.3, 427b11-12).

marks of intentionality, however, is that intentional states can be about objects that don't exist and hence could not have caused them—intentional states can be false. Aristotle's account of contentful perceptional states thus fails to give an explanation of intentional states since it does not explain their capacity for error. He remedies this shortcoming, however, by introducing *phantasia*, a process typically translated as "imagination," or "appearing."

Phantasia, and the *phantasmata*—"images" or "appearances"—that it produces, rests upon the simple analogue process of perception:

> "...phantasia seems to be a sort of change that does not occur without sensation, but belongs to perceivers and *is about what the sensation is about*, and [since]...it is possible for a change to occur due to the functioning of sensation and *this change is necessarily similar to the sensation*, [it follows that] this change could not occur without sensation or without belonging to perceivers, and *its possessor can both do and undergo may things in accordance with it* and *it can be both true and false*."[45]

To use Caston's analogy, a *phantasma* is like an 'echo' of the sensory perception that caused it. This foundational causal link with perception lends *phantasmata* their basic character. Because they occur "due to the functioning of sensation," and thus *normally* have the same content as sensations, *phantasmata* are ensured of having causal powers similar to the sensations that produce them. Like sensations, for example, *phantasmata* can function in a subject without being asserted or accepted by that subject, an act that would help graduate a *phantasma* to the status of belief.[46] They can thus play a role in the production of animal behaviour, and in some special forms of human behaviour that are unguided by thought: "...it is on account of these [*phantasmata*] persisting and being

---

[45] *On the Soul*, III.iii., 428b10-17. In Caston 1998, p. 273.

[46] *On The Soul*, 3.8, 432a10-11; *Movement of Animals* 6, 700b16-17, 701a4-6. On the distinction between

like sensations that animals do many things accordingly, some who lack intellect (namely beasts) and others due to the intellect's being clouded over on occasion by passion, illness, or sleep (namely humans)."[47] More particularly, Aristotle claims that *phantasmata* can sometimes play a role in the production of emotions, especially fear and shame.[48] *Unlike* sensations, however, *phantasmata* can sometimes change in ways that radically alter their original causal powers, just as an echo can change so as to sound different than the noise that originally produced it. Caston offers Aristotle's example of one's perceptions of a salamander that are transformed, either through sickness or alcohol, into the vision of a fire-breathing dragon. Unlike sensations, then, phantasmata can *diverge* from their causal source and as a result gain new causal powers, e.g., the power to lead one to think "Lo! A Dragon!" instead of "Lo! A Salamander!". This divergence, in turn, opens up the possibility of error, since the content of a *phantasma* is determined by its causal powers. Caston notes:

> A dragon can't be a causal ancestor of my dream—dragons don't exist. But my phantasmata have the ability to effect my central organ the way it *would* be affected *were* I to see such a dragon. The causal history of phantasmata is thus not relevant to their content except *per accidens*. At most, it can explain why a phantasma has the particular causal powers it happens to have. But its content is solely a function of the powers it actually does have at a given moment, *however it came by them*....The content of phantasmata can thus diverge completely from

---

phantasmata and belief, see *On The Soul*, 428a18-24.

[47] *On The Soul* 3.3, 428b30-429a9. Nussbaum (1994, pp. 291-92) points out that Aristotle's claim that humans could in some circumstances be moved to action by phantasmata alone was radically extended by the Skeptics, who argued that one could withhold *all* acts of assent and thus behave in ways based entirely upon *phantasmata*, withholding assent completely. She quotes Sextus Empiricus' *Outlines of Pyrronhism*: "Clinging to appearances [*phantasmata*], then, we live without belief, according to the practices of life, since we cannot be altogether inactive."

[48] Aristotle specifically mentions *phantasmata* in his definitions of fear (*Rhetoric* 1382a21-3, 1382a28-30, 1383a17) and shame (1384a23). However, in *On The Soul* (427b21-24), Aristotle notes that one can imagine a fearful or threatening scene and not become frightened of it, whereas *thinking* that a situation is dangerous will always immediately lead to fear.

> their causal ancestry and, more generally, from what is actually the case—*they can be false.*[49]

With this account of the intentionality of *phantasmata* in hand, the final step in a complete theory of intentionality is to explain the process whereby more complex mental states like belief are grounded in *phantasmata*.

Caston argues that Aristotle conceives of this process as an act of the faculty of understanding in which *phantasmata* are 'transduced' into *concepts* by the understanding's actively *ignoring* certain aspects of a *phantasma's* content.[50] For present purposes, however, this particular issue isn't really important, since an account of *phantasmata* alone serves to illustrate Aristotle's basic theory of intentionality. One aspect of this issue, however, is relevant here. Insofar as *phantasmata* stem from perception, and so retain the character of perceptions, it follows for Aristotle that they are in some sense 'non-propositional.' Precisely how this common claim is to be cashed out is unclear.[51] Phantasmata are not, for example, mental 'pictures,' since they follow on perceptions produced by all sensory modalities. At the very least, however, it is clear that phantasmata are importantly different than even simple concepts and beliefs, and part of this difference will be spelled out by contrasting *phantasmata* with purely linguaform

---

[49] Caston 1998, p. 275.

[50] "To arrive at this higher level of representation requires a different power in Aristotle's opinion, the power of conception or understanding (νοϛ), which grasps part of a phantasma's content to the exclusion of others in a new mode—— again, a form of transduction. It is unlikely it does this in the way imagined by some later Aristotelians, by literally stripping away matter from the phantasma and leaving the bare concept. To the extent that Aristotle himself says anything about the subject, he seems to have in mind, not so much the production of a separate entity, as a different way of handling the phantasma, by *ignoring* certain features (*On Memory* 1, 449b30-450a14). Different phantasmata, that is, *can be treated as equivalent*, insofar as they each have a certain part of their content in common; and that aspect of a phantasma which allows it to be treated in this way would be a concept or νοημα (as distinct from the *object* of thought or νοητόν)" (Caston 1998, p.225).

[51] As Nussbaum (1994, p. 85) points out, since "the [phantasma]of the sun as a foot wide involves, at the very least, *combination* or *predication*...[it] is a little hard to see where to draw the line between this and

representations.[52] It follows then that to the extent that *phantasmata* play a role in the production of a given emotion, that emotion will fit poorly into the analytical framework employed by Aristotle to represent the cognitive structure of emotions. This framework, recall, depends upon the existence of a logical relationship between the cognition *token* that produces an emotion and the cognition *type* that is helps define that emotion as an instance of a particular *emotion* type. If, however, an emotion is caused by a *phantasma*, it is difficult to see how to align that *phantasma* token with *any* cognition type, since those types are necessarily linguistically individuated and *phantasmata* are, in some sense, non-linguaform.

While this is an important issue, I won't pursue it here, as I will discuss its modern counterpart in subsequent chapters. Moreover, it is not clear how important Aristotle actually thought *phantasmata* were in the production of emotion. Given the complexity of the cognitions he typically held to produce emotion it is unlikely that he thought *phantasmata* were a pervasive cause. Even so, I would suggest that it still remains an important fact about Aristotle's theory of emotion that he at least recognises the possibility of an emotion's being grounded in cognitive states less robust and complex than belief.

Finally, however important *phantasmata* actually are to Aristotle's theory of emotion, they at least help explain how emotions are capable of intentionality, since an emotion *qua* unified psychological state will *derive* its intentionality from that of the cognitive state in which it is grounded. And while this state might typically be a

---

the "propositional"."

complex, linguaform belief, on Aristotle's account all such states will ultimately derive *their* intentionality from that of *phantasmata*.

At this point I want to conclude my discussion of Aristotle's theory of emotion with a more general observation, for to focus wholly upon his conception of how emotions are structured and individuated would be to miss an important aspect of his overall theory that lies hidden in the details. Aristotle's emotional subjects move in an essentially *social* world. They are not abstract cognizers but real people of determinate social standing and character who live and move in a complex system of social relationships. They are old or young, rich or poor, supplicant or benefactor, ruler or subject, dear friend, minor acquaintance or utter stranger, enemy or ally in war, superior in birth or inferior in power.[53] In Aristotle's analysis we are repeatedly shown how these social roles and relations figure directly in the modulation of our emotions: "Anger is felt towards friends, in a greater degree than towards such as are not friends....Men feel [anger] also in a greater degree towards persons of no account, should they slight them."[54] Pity is felt with particular ease "towards [one's] equals, whether in age, in temper, in habits, in rank, or in family; for in all these relations the evil is seen with greater clearness as possible to befall also one's self."[55] Equality in "circumstances of birth, connections, age, habits, character, and property" is likewise the prerequisite for envy, which Aristotle defines as a pain occasioned by the unwarranted good fortunes of

---

[52] "...[*phantasia*]is different from either perceiving or *discursive thinking*, though it is not found without sensation, or judgement without it" (*On The Soul*, 427b14-15).
[53] They are even living or dead (*Rhetoric*, II.iii.xv).
[54] *Rhetoric*,II.ii.xiv.
[55] *Rhetoric*, II.viii.xiii.

those we hold as equals in these things.[56] Regarding shame, a sort of pain attendant upon the loss of character, Aristotle notes that "people are not at all sensible of shame before those whose opinions, in regard to their justness, they hold cheap...no one feels shame before children and brutes."[57]

What is implied in Aristotle's detailed accounting of the impact of the social world on our emotions is that the cognitive aspect of our emotions is often significantly more complex than it appears in his simple definitions. Pity, for example, involves more than the basic recognition of an evil about to befall another; it further involves the recognition of that person as an equal in some particular aspect. Similarly, shame requires more than the basic recognition that we have suffered a loss of character; it further involves the positive judgement of another's opinions of justness. Of note here is that these supplemental judgements are generally all of a single nature: they are evaluations. Here I use evaluation in the sense of an evaluation of x being a judgement of the *worth* of x.[58] In envy, my recognition of you as an equal involves my judging particular aspects of you— e.g., your birth, social standing, or general moral character— to be sufficiently worthy, where that judgement proceeds against the benchmark of my own circumstances and standing. Similarly, my shame before you depends upon a judgement that your opinions of justness are of a worth that merit my discomfort in front of you. Continuing down this avenue, it is further clear that the evaluations of worth underlying our various emotions themselves rely upon a widening range of judgements,

---

[56] *Rhetoric,* II.x.i.
[57] *Rhetoric*, II.vi.xxiii.
[58] As opposed to, say, judgments of the identity of x, or judgments that x possesses a particular physical quality.

beliefs, and evaluations. Some of these cognitions involve simple factual judgements—e.g., about the identity of one's parents—while others are themselves further judgements of worth.

I won't elaborate further on this point. My goal here is not to provide an analysis of all cognitions that figure in the production and identity of our particular emotions, nor is it Aristotle's concern in the *Rhetoric*. My concern is instead to give some idea of the cognitive complexity that Aristotle saw as both underlying and partially constituting our emotional lives, and moreover, to emphasise the *social* nature of these cognitions.

In sum, then, the picture of emotions that emerges from the *Rhetoric* can be stated as follows:

1. Emotions are highly structured intentional states—consisting of cognitive, motivational, and hedonic elements—that depend for their existence upon a broad base of background beliefs and judgements about oneself and their position in an essentially social world. Aristotle thus complicates the simple Platonic conception of a monolithic emotional 'faculty.'
2. Emotion types are defined by the formal structure of their cognitive component in that each emotion type is related to a unique judgement type. Particular occurent emotions are individuated according to which judgement type their actual eliciting judgement falls under.
3. Emotions have an ambiguous relationship to reason and argument. They can both influence, and be influenced by, our considered judgements. The mechanism by which this happens, however, is left unexplained by Aristotle.

Descartes: Perception, Representation, and the Function of Emotion

Les Passions de l'Ame occupies a curious place in both Descartes' own oeuvre and in the history of philosophical and psychological treatments of emotion.[59] Given its date (1649) it is surprisingly modern in many of its technical points; given its author's arch-rationalism it is surprisingly sympathetic in its treatment of the worth and rationality of emotion. Finally, and most intriguingly, Descartes' treatment of emotion's ambiguous positioning between the categories of activity and passivity often edges toward a dissolution of the radical mind-body dualism that underlies the entire theory.

The tone of the work is strongly scientific. In the preface Descartes announces his "intention to explain the passions only as a natural philosopher, and not as a rhetorician or even as a moral philosopher."[60] Here Descartes explicitly sets himself against both Aristotle and the long philosophical tradition of approaching emotion as an object of moral study. Emotions, for Descartes, are not the lamentable weaknesses of an irrational will or dysfunctional soul but are rather the inescapable and necessary accoutrements of the human body.

More specifically, emotions are for Descartes a special class of perceptions and so deserve the general name 'passions' in that they belong to a class whose defining feature is *passivity*—our perceptions are visited upon us by causes in the world external to our

---

[59] All references to Descartes in this section are to Descartes 1985. References to *The Passions of the Soul* (*PS*) are to section number; references to *Optics* (*Opt.*) are to discourse and paragraph number; references to *Treatise on Man* (*TM*) are to paragraph number; references to *Treatise on Light* (*TL*) are to chapter and paragraph number.

[60] *PS*, preface.

body and soul and so are largely beyond our control.[61] What is it, however, that distinguishes our passions *proper* from our simpler perceptions? Descartes' answer here rests on the notion of referral. Our passions proper—our "feelings of joy, anger and the like"—are "the perceptions we refer only to the soul."[62] Passions qua perceptions thus differ from perceptions of colour and shape, and perceptions of pain and hunger, in that we refer the former type to objects external to the body and the latter to our body or its particular parts. What is it to 'refer' a perception to some object? This is a crucial question for Descartes and he is unfortunately obscure. In part, referral involves the subject making a judgement of cause. Regarding our perceptions of the external world Descartes notes that "we refer these sensations to the subjects we suppose to be their causes in such a way that we think that we see the torch itself and hear the bell, and not that we have sensory perception merely of movements coming from these objects."[63] So we refer our perceptions of external objects to those objects by judging that they are the causes; we think we see the wax itself as opposed to thinking we are merely experiencing rays of light reflected from its surface. Referral also seems to involve, though, what might loosely be called 'judgements of location.' Perceptions that we refer to the body—such as hunger and thirst—include any states "we feel as being in our limbs, and not as being in objects outside us."[64] Similarly, the perceptions "we refer only to the soul are those whose effects we feel as being in the soul itself, and for which we do not normally

---

[61] Descartes uses "perception" to signify "all the thoughts which are not actions of the soul or volitions" (*PS* §28). There is, however, a subclass of perceptions that are inseparable from those volitions and hence are *active* perceptions, namely, those perceptions of the soul's acts of volition.
[62] *PS* §25.
[63] *PS* §23.
[64] *PS* §24.

know any proximate cause to which we can refer them."[65] Here referral involves both forms of judgement: we both feel the sensation in a discrete location within us, and given that we lack an apparent external cause of the perception, we judge its cause to lie in our soul.[66]

Unfortunately, as a principle by which to individuate our emotions from the larger class of perceptions, the notion of referral is a weak one for the simple reason that it clearly is not the case that we feel emotions only 'in the soul.' Descartes himself says as much in a passage where he seeks to dispel the popular myth that the *heart* is the seat of the passions. He explains there that the source of this confusion lies in the fact that "the passions make us feel some change in the heart."[67] This physical change is subsequently explained by the presence of a nervous connection from the brain to the heart, but whatever the explanation, Descartes' notion of referral is in trouble. If 'referring' a passion involves having some sense of its location then we clearly *feel* our emotions throughout our entire body. Descartes theory of the emotions in fact offers a detailed cataloguing of the physical effects of our emotions, and moreover explains *why* we feel our emotions in the places and manner in which we do. Does Descartes then have no principled way of individuating the class of passions proper? I would argue that he in fact does, and that his answer to this question marks him as a thoroughly modern emotional theorist.

---

[65] *PS* §25.

[66] Descartes suggests that "it is even better to call them 'emotions' of the soul...because of all the kinds of thought which the soul may have, there are none that agitate and disturb it so strongly as the passions"(*PS* §28). Here Descartes is relying upon "emotion" in its original sense—to move.

[67] *PS* §33. See also §31.

As noted above, it is the passivity of the passions proper that leads Descartes to see emotion as a form of perception. The passions proper, however, share more than just passivity with our common perceptions of the world. Specifically, common perceptions and the passions proper both involve *representation*. Perception is for Descartes that process whereby the world operates upon the body—hence perception's passivity—to create representations in the brain that are subsequently considered by the soul. What sets the passions off as a unique class of perceptions is that they involve representations of a unique type. Exactly *what* representation in general involves, and *how* passion's representations are unique—how they differ from the representations involved in our perception of shapes and such—becomes clear when we look at the details of Descartes' theory of perception.

As developed in the *Optics* and the *Treatise on Man* Descartes' theory of perception is straightforwardly mechanistic. The transfer of perceptual information from objects to perceiver is wholly explained by the interaction of minute physical bodies. In the first step of visual perception, for example, an image of the object being perceived is formed on the back of the eye "through the medium of...intervening transparent bodies."[68] At this point the image is a "perfect likeness" of the perceived object, produced in essentially the same way as the images of a *camera obscura*.[69] As this image falls upon the retina, tiny tubes—the optic nerves—that connect the eyes to the brain are opened, small particles in the tube are set in motion, and in this way "a corresponding

---

[68] *PS* §13.
[69] *Opt.*, V, 130.

figure [is] traced on the internal surface of the brain."[70] There is a similar transaction between the internal surface of the brain and the surface of the pineal gland. The end result of this transaction is that a "figure is traced on the surface of the gland."[71]

We have then a number of 'figures' or images occurring at various points in the perceptual chain: images on the retina, figures on the brain's internal surface, and figures traced on the surface of the pineal gland. These last figures, though, are unique in that they are the only ones that might rightly be called "ideas."[72] Being physical states, however, they cannot truly be Cartesian ideas. It is more precise to say, as Descartes does, that "it is only the latter figures which should be taken to be the forms or images which the rational soul united to this machine [i.e., our body] will consider directly when it imagines some object or perceives it by the senses."[73] So the final *physical* link in perception's causal chain is a figure traced on the surface of the pineal gland. This figure, in turn, is taken to "represent to the soul" the various physical properties of the object perceived. In relation to emotion, these figures, which there play a central causal role, are of a unique type.

Consider, for example, Descartes' account of the definition and causes of wonder, the first of what he sees as the six basic emotions:

> Wonder is a sudden surprise of the soul which brings it to consider with attention the objects that seem to it unusual and extraordinary. It has two causes: first, an *impression in the brain, which represents the object as something unusual* and consequently worthy of special consideration; and secondly, a movement of the spirits, which the impression disposes both to flow with great force to the place in the brain where it is located so as to strengthen and preserve it there, and also to

---

[70] *TM* 175.
[71] *TM* 176.
[72] *TM* 177.
[73] *TM* 177.

> pass into the muscles which serve to keep the sense organs fixed in the same orientation so that they will continue to maintain the impression in the way in which they formed it.[74]

The first point to notice here is Descartes' reference to an "impression in the brain." Here "impression" is just a synonym for the pineal figures that are the end physical result of the mechanical process of perception. Notice, though, the complex representational features that wonder's defining pineal figure is claimed to possess: it represents an object to the soul "as something unusual." Similarly cognitively complex representations are likewise definitional of the other basic emotions. Love involves representations of the beloved object "as agreeable"; hatred, conversely, is directed toward objects "which are presented [to the soul] as harmful."[75] Joy, another basic emotion, occurs when the soul "enjoys a good which impressions in the brain represent to it as its own."[76] Sadness, conversely, occurs when the soul experiences "an evil or deficiency which impressions in the brain represent to it as its own."[77] The point of this cataloguing is to give some idea of the complexity that Descartes ascribes to particular alterations in the surface of the pineal gland. Abstracting slightly from the particulars we can characterise all of these representations as roughly being of a single sort, namely, representations of complex 'self-relational' properties. That is, all of these representations portray various basic ways in which the objects being perceived might relate to the 'self' perceiving them. These various ways of relating divide along two dimensions: objects might be related to the perceiving subject as (1) harmful/beneficial, or (2) belonging/not-belonging. These

---

[74] *PS* §70.
[75] *PS* §79.
[76] *PS* §91.
[77] *PS* §92.

two basic dimensions need only to be supplemented with the concept of time, in the case of desire, to account for all of the self-relational properties that are represented in the six basic emotions.[78]

That the representations uniquely involved in emotion should all be of this particular type is to be expected, given Descartes account of emotion's main function. Notice in the definition of wonder how the perceptual-causal chain that produces the pineal representation of an object as being novel and worthy of consideration also directly moves the body to actions that ensure continued sensory contact with that object. This particular account is just a special case of the general functional principle that the physiological changes that cause the representing pineal figures also act directly upon the body in ways germane to the situation at hand:

> The function of all the passions consists solely in this, that they dispose our soul to want the things which nature deems useful for us, and to persist in this volition; and the same agitation of the spirits which normally causes the passions also disposes the body to make movements which help us to attain these things.[79]

The *telos* of emotion is preservation, and it proceeds to this end by producing an alignment of thought and action that works to place the subject in the most beneficial stance to a particular situation. Given this conception of emotion's function, then, it is to be expected that the representations involved in emotion will generally represent the particular self-relational aspects of a situation that they do. If the function of emotion is essentially preservative, then *any representations particular to emotion should be of those aspects of our relation to the world that are significant for our well-being.* I will

---

[78] Descartes defines desire as "an agitation of the soul caused by the spirits, which disposes the soul to wish, in the future, for the things it represents to itself as agreeable" (*PS* §86). Considered abstractly, desire is thus a form of love in that it involves representations of the desired object as agreeable.
[79] *PS* §52.

expand on this point below, but before doing so I want to turn now to a fundamental question that this account of emotion raises. Given the centrality of representation to emotion, and the representational complexity that is ascribed to the figures traced on the pineal gland, we must ask: *how* do these figures, which are essentially minute alterations in the physical structure of the brain, represent complex properties to the soul? More simply, how do they represent at all? The short and striking answer here is that, strictly speaking, these figures do not in fact represent anything *to* the soul. The longer answer, developed below, is that Descartes is implicitly relying here upon a causal theory of representation, i.e., a theory in which all talk of representation ultimately factors out into strictly causal terms.

In his discussion of perception in the *Optics* Descartes makes the general cautionary point that sensory perception does not require the soul to "contemplate certain images transmitted by objects to the brain."[80] If, however, one wishes not to step too far outside of tradition it is acceptable to speak of the soul's contemplating images so long as we conceive the nature of these images in an entirely different manner from that of the philosophers.[81] Descartes point here is that it is wrong to think that images contemplated by the soul must in any way *resemble* the objects that they represent: "We should recall that our mind can be stimulated by many things other than images—by signs and words, for example, which in no way resemble the things they signify."[82] This claim is supported with a number of examples. Despite consisting only of blotches of ink on paper, etchings are capable of representing not only complex physical structures like

[80] *Opt.*, IV, 112.
[81] *Opt.*, IV, 113.

trees, towns and people, but can also "make us think of countless different qualities in these objects, [though] it is only in respect of shape that there is any real resemblance."[83] Even more strikingly, words are capable of representing an infinite variety of things even though they bear no resemblance to the things they signify.[84]

So Descartes allows talk of images properly construed. Of course, to point out that representation does not depend on resemblance is not to deny that the soul perceives representations that are physically instantiated on the pineal gland. Two fundamental problems remain. First, the question of how physical structures in the brain represent has only become more pressing: if not by resemblance then how? Secondly, Descartes has not yet escaped the circularity of accounting for external perception in terms of the soul's *inner* perception of figures in the brain.

The strategy that Descartes adopts here is to rephrase the pressing questions and adopt a new theoretical language. He does this by abandoning talk of resemblance and representation, and focusing instead upon the *causal role* of the figures traced on the pineal gland: "The problem is to know simply how [these figures] can enable the soul to have sensory perceptions of all the various qualities of the objects to which they correspond—not to know how they can resemble these objects."[85] Descartes' approach thus shifts from wondering how the brain can represent objects to the soul to searching for a more precise account of the causal link between the physical properties of the pineal figures and the experiencing of sensations of colour, shape, and size.

---

[82] *Opt.*, IV, 112.

[83] *Opt.*, IV, 112. In this passage, and elsewhere (see paragraph 135) Descartes explicitly notes the difficulty in speaking of perception in non-representational terms. He points out in these passages the grip that our common understanding of pictures has had on philosophical conceptions of perception.

Descartes proceeds in this search by first insisting that the pineal figures are to be properly understood *just as* the final physical link in the perceptual-causal chain. His account of perception thus does not require that figures in the brain work as representations in the sense that their functioning does not depend upon reception in the soul by some further perceiver.[86] Rather, he claims now that they function in a directly causal way upon the soul and that subsequently no inner *perceiver* is required. These points come together explicitly in the following passage. Here, Descartes initially notes that in the first stages of perception we can rightfully speak of images as representing by resemblance the objects perceived, as when the image of some object first falls upon the retina. He cautions us again, however, that such resemblance plays no *functional* role in the soul's experience of the objects perceived:

> Now, when this picture [formed on the retina] thus passes to the inside of our head, it still bears some resemblance to the objects from which it proceeds. As I have amply shown already, however, we must not think that it is by means of this resemblance that the picture causes our sensory perception of these objects—as if there were yet other eyes within our brain with which we could perceive it. Instead we must hold that it is the movements composing this picture which, acting directly upon our soul in so far as it is united to our body, are ordained by nature to make it have such sensations.[87]

Talk of inner perception, or 'representations to the soul' thus factors out into strictly causal terms. Representations *qua* physical pineal figures do not represent by virtue of being inwardly perceived pictures of the objects that are outwardly perceived by the senses; they rather 'represent' by virtue of their ability to cause us to experience

---

[84] *TL* I, 4.
[85] *Opt.*, IV, 114.
[86] "For an image to work as an image there must be a person (or an analogue of a person) to see or observe it, to recognize or ascertain the qualities in virtue of which it is an image of something"(Dennett, 1969, p. 134).
[87] *Opt.*, VI, 130.

immaterial sensations. My claim here, in short, is that Descartes does not think of the pineal figures that are the final physical link in the perceptual-causal chain as being representations in any but the most attenuated sense. Or, more precisely, he may allow that they are representations of some sort but argue that nothing that is essential to their being representations in any *traditional* sense of the term need necessarily figure into their *causal* role in determining the experiences the soul has during perception. At this point, though, we are immediately led to ask what features of these figures—if not their resemblance to the objects they represent— *do* function causally in the production of perceptual and emotional experience.

Descartes' answer here rests upon what he takes to be the fundamental principle underlying his entire theory of emotion. It is for Descartes a brute and *prima facie* mysterious fact about human physiology and psychology that "nature or habit has joined certain movements of the [pineal] gland to certain thoughts."[88] The movements referred to here are just those particular structural changes on the pineal surface that are the final physical link in the perceptual-causal chain. To say that nature or habit 'joins' particular thoughts to particular movements of the gland is just to say that nature compels particular

---

[88] *PS* §44. While the connection between particular thoughts and movements of the gland (and hence to emotions) is originally ordained by nature it is not immutable. We can alter this original linkage through indirect exploitation. To explain how we might do so Descartes draws an analogy between our passions and the dilation of our pupils (*PS* §44). We cannot consciously control the dilation of our pupil because nature has not established a link between the pineal movements that control pupilary dilation and the volition to do so. It has, however, established a link between these movements and the volition to view distant objects: when we wish to view a faraway and act upon it, by casting our gaze into the distance, our pupils naturally dilate. So, we *can* consciously dilate our pupils, though only indirectly, by exploiting the original connection established by nature: if we want to dilate our pupils we need only look at faraway objects. The same holds true for our passions. Like the dilation of our pupils, "our passions, too, cannot be directly aroused or suppressed by the action of our will, but only indirectly through the representation of things which are usually joined with the passions we wish to have and opposed to the passions we wish to reject" (*PS* §45). We can thus overcome our fear, Descartes suggests, by bringing to mind the thoughts that nature has naturally associated with its opposites, bravery and boldness.

movements to cause particular thoughts.[89] Clearly, though, positing such a connection based on nature's fiat is hardly explanatory. It remains to be explained *why* nature or habit connects one particular set of thoughts and desires to a particular figure rather than another. Are these connections governed by general rules? Do they depend upon features intrinsic to the movements of the gland? To the content of the relevant thoughts? Descartes does have a detailed answer available to this general line of questioning, one that emerges in his discussion of the physiological foundations of emotion.

Descartes claims that our various basic emotions—wonder, love, hate, joy, sadness and desire—have their origins in correspondingly distinct types of physiological changes. The general principle he proposes runs as follows. In the course of an individual's emotional development there will occur at some point a set of physiological changes resulting from contact with a highly significant object or situation. These changes cause an initial basic passion that is henceforth associated with the physiological effects of the original set of bodily changes. Consider, for example, Descartes' explanation of the origins of love:

> It seems to me that when our soul began to be joined to our body, its first passions must have arisen on some occasion when the blood, or some other juice entering the heart, was a more suitable fuel than usual for maintaining the heat which is the principle of life. This caused the soul to join itself willingly to that fuel, i.e. to love it.[90]

Love is thus first caused by some beneficial physiological change. Similarly, an individual's first experience of sadness arises when "it has happened that the body has

---

[89] Admittedly, this account leaves mysterious the particular causal mechanism involved at the point of the body's contact with the soul; dualism's fatal flaw is not overcome here. Setting this problem aside, though, we can still ask what more Descartes can say about the nature of these connections.

[90] *PS* §107.

lacked nourishment."[91] We have, then, a simple account of the origin of particular emotions in bodily changes of a particular type that themselves result from contact with a particularly important type of situation. The question, though, is *why* love, rather than hate or sadness? Why is love's definitive constellation of thoughts and desires, rather than some other, initially connected to these first beneficial physiological changes?[92]

Descartes' answer rests upon his claim that the function of emotion is to "move the soul to consent and contribute to actions which may serve to preserve the body or render it in some way more perfect."[93] If we take this as axiomatic—that the essential function of emotion is preservation—it immediately follows that the general rule governing which cognitive sets are to be joined to a particular 'representing' pineal figure will be something like: the particular thoughts or volitions initially caused by physiological changes, and thereafter associated with those changes, are to be exactly those that will serve to impel the emotional agent to place themselves in the most beneficial relation to the situation eliciting the emotion. Thus given emotion's preservative *telos* we should *expect* that love—the essence of which is the desire to join oneself to that which is conceived of as beneficial—will naturally be joined to those physiological changes which are beneficial. Similarly, it follows that hatred—the essence of which is the desire to remove oneself from whatever is harmful—will naturally be joined to physiological changes that are damaging or dangerous to the body.

---

[91] *PS* §110.

[92] Strictly speaking, of course, an emotion's definitive cognitive set is not connected to the physiological changes considered grossly, but only to those changes in the pineal gland which arise from those more basic physiological changes.

[93] *PS* §137.

To link this point with an earlier question about representation, we now have a clearer idea of exactly which features of the representing pineal figures will play a direct causal role in eliciting particular emotions. In short, it will be those features of the figure that are causally linked to the emotionally germane features of the situation eliciting the given emotion. By 'emotionally germane features' I mean those features of the relevant situation that determine how we are to situate it along the various self-relational dimensions. More simply, the features of the representing figure that determine which thoughts should be connected to that figure will be those features that are directly caused by the features of the situation which make it harmful or beneficial for us, and that show that it does or does not belong to us. Descartes thus notes that "the objects which stimulate the senses do not excite different passions in us because of differences in the objects, but only because of the various ways in which they may harm or benefit us, or in general have importance for us."[94] In sum, then, Descartes does provide a general explanatory principle that can explain *why* certain passions/thoughts/desires are connected with particular bodily changes.

While there is a great deal more of interest in Descartes' account of the emotions, I want to close now by taking from Descartes the following themes:

1. Our basic emotions function to effect a beneficial alignment of body and mind in response to distinct types of highly significant situations. Wonder, for example, promotes a set of physical and mental responses to "unusual and extraordinary objects" that work together to place the subject in the best possible relation to those objects. This conception prefigures the modern view that emotions are "tools to

[94] *PS* §52.

promote psychobehavioral coherence" in the face of "categories of life-challenging events."[95] I will say more about this notion in later chapters.

2. This understanding of emotion's function allows Descartes to explain *why* our various emotions have the 'cognitive structure' they do, a question left unaddressed by Aristotle. In each particular case the original linkage of bodily and mental response is determined by some historical occurrence in the course of an individual's emotional development. However, the structure of this original connection—i.e., the content of the thoughts and volitions associated with the initial bodily changes—does not occur by chance but is rather guided by the preservative *telos* of emotion. Of course, Descartes can't ultimately explain why we should have designed into us states that perform this function; presumably he must say they were placed there by God. A full answer to this deeper question will have to wait for the appearance of Darwin, to whom I now turn. Short of this, however, Descartes does provide a general principle that explains *why* particular emotions are associated with the thoughts and desires that they are.

Darwin: Emotion, Expression, and Evolution

Darwin's *The Expression of the Emotions in Man and Animals* is a seminal work in emotion theory. First published in 1872, Darwin intended his painstakingly detailed study of human and animal expressive behaviour to form a central plank in his argument

[95] Panksepp 1998, pp.39, 55.

for the evolution of species through natural selection. Expression's value in this regard rested upon two points.

Darwin reasoned that the universality of human emotional expression would be strong proof for the claim that the various human races had evolved from a single common ancestor. To find that sophisticated Europeans and the wildest Aborigines all similarly wrinkled their brows when perplexed, blushed with shame, and opened their eyes and mouths wide when astonished, would imply that such behaviour is not conventional: "Whenever the same movements of the features or body express the same emotions in several distinct races of man, we may infer...that such expressions are true ones—that is, are innate or instinctive."[96] Such shared innate behaviour, Darwin argued, points in turn to a common ancestor for the varied human races.

Even more central to Darwin's case for evolution, though, are the expressive similarities that exist between humans and animals. As Paul Ekman points out, proof of a common ancestry for the human races need not convince a creationist of evolution's truth since they could simply argue that this is just proof we have all descended from Adam.[97] It was thus important for Darwin to find expressive behaviour shared *across species*. Universality of this type made a stronger case for evolution: "The community of certain expressions in distinct though allied species, as in the movements of the same facial muscles during laughter by man and by various monkeys, is rendered somewhat more intelligible if we believe in their descent from a common progenitor."[98] To convincingly make his case for evolution, however, Darwin had to do more than simply point to cross-

---

[96] *The Expression of the Emotions in Man and Animals*, p. 22.
[97] Ibid., p. xxvii.

species similarities in the expression of emotion. For such universality to be convincing proof of evolution, Darwin had to show that expressions common across species had evolved largely for the same reasons and hence were explicable by the same principles.[99] Darwin's fundamental approach to expression, therefore, was to search for explanations of shared expressive behaviour that tied humans to the 'lower' animals, rather than setting them off as unique. Not surprisingly, this placed Darwin squarely in opposition to the theories of expression dominant in his time.

When *The Expression of the Emotions* was published in 1872 a number of important works on expression already existed. Chief among these was Charles Bell's *Anatomy and Philosophy of Expression*. Darwin credited Bell with having "laid the foundation" for the scientific study of expressive behaviour, and made extensive use of Bell's work in descriptive anatomy and of his detailed observations of human emotional expression.[100] Despite such praise, however, Darwin saw Bell's approach to expression as fatally flawed. Bell—who was no evolutionist—argued for the widely held view that humans had been endowed by God with a unique musculature intended for the sole purpose of expression. To Bell's mind the human capacity for expression thus served as a marker of human distinctness. Not surprisingly, this was anathema to Darwin.

In his introduction to *The Expression of the Emotions* Darwin first argued against Bell by pointing out that the musculature involved in human expression was not as unique as Bell claimed. Anthropoid apes and humans, for example, have essentially the same facial musculature. On Bell's view this fact leads to the decidedly odd conclusion

---

[98] Ibid., p. 19.
[99] Ibid., p. 25.

that humans and apes were endowed with the same musculature for entirely different reasons. The other option, equally unsavoury, would be to " admit that monkeys have been endowed with special muscles solely for exhibiting their grimaces."[101] Both options were equally unacceptable to Darwin and this introductory argument is meant to illustrate the basic weakness in Bell's position.

Darwin, however, carried his argument with Bell far beyond the introduction; it shaped and informed his entire approach to the subject, most importantly influencing his choice of explanandum.[102] As noted above, Darwin focused on explaining behaviours shared across races and species since such universality implied descent from a common animal ancestor. He focused more specifically, though, on universal simple and *involuntary* behaviours. More complex, *conventional* forms of expressive behaviour—such as ritualised gesture and linguistic description—were purposefully left unexplored.[103] The reason here is straightforward.

In general, the universality of any particular form of behaviour is *not*, on its own, proof that the behaviour has a genetic basis. Dennett provides an instructive example. Humans everywhere throw their spears pointy end first yet the ubiquity of this behaviour

---

[100] Ibid., p. 7.
[101] Ibid., p. 17.
[102] Disproving Bell's claims about the uniqueness and purpose of human 'expressive' musculature was a major factor in moving Darwin to write *The Expression of the Emotions*. In March 1867 Darwin wrote to Alfred Wallace: "I want, anyhow, to upset Sir C. Bell's view...that certain muscles have been given to man solely that he may reveal to other men his feelings"(quoted in Ekman's comments, p. 8). Writing on this point in his *Autobiography* Darwin recalled: "During the summer of the following year, 1840, I read Sir C. Bell's admirable work on Expression, and this greatly increased the interest which I felt in the subject, though I could not at all agree with his belief that various muscles had been specially created for the sake of expression" (ibid.).
[103] Darwin explicitly claimed that our involuntary, innate behaviours "alone deserve to rank as true expressions" (p. 55).

is clearly not proof of a 'pointy end first' gene.[104] Throwing a spear in this fashion is simply the most reasonable thing to do; it is the most rational solution to a particular problem, and with humans being the rational creatures they are, we should expect such 'pointy-end-first' behaviour to be universal. However, the *less* rational a piece of behaviour is, and the less it is under conscious control, the *more* its universality can be considered as proof for a genetic foundation, since in these cases such appeals to rationality lose explanatory force. Darwin's case against Bell could thus be made most strongly by explaining the universality of *involuntary* expressive behaviours.

Beyond his particular disagreements with Bell, however, Darwin saw Bell's difficulties as a special case of a more fundamental error shared by virtually every theory of expression that had preceded his own. Darwin argued that any study of expression not grounded in evolutionary principles was bound to lack true explanatory force: "No doubt as long as man and all other animals are viewed as independent creations, an effectual stop is put to our natural desire to investigate as far as possible the causes of expression. By this doctrine [of Creation], anything and everything can be equally well explained."[105] Darwin's point here is that the invocation of God's fiat serves to explain everything and thus explains nothing. Regarding previous theories of expression in particular, Darwin argued that appeals to a Creator had put an effective halt to the all important question of *why* we express our emotions in the ways we do. Why, for example, do we furrow our brows when distressed? Blush when embarrassed? Tremble when frightened? And why do humans share so many forms of expression with other species? Creationist accounts,

---

[104] Dennett 1995, pp. 486-7.
[105] Ibid., p. 19.

Darwin claimed, had effectively ruled out meaningful answers to these questions, and to prove this point he quoted at length "specimens of the surprising nonsense" written on the subject.[106] Expression was thus fertile ground for proving the explicative force of evolutionary theory.

Darwin proceeded to this end by arguing that most *involuntary* expressive behaviour could be explained by one of three general principles: the *principle of serviceable associated habits*, the *principle of antithesis*, and the *principle of direct action of the nervous system*. Of the three, the first is most important. Darwin explains it as follows:

> Certain complex actions are of direct or indirect service under certain states of the mind, in order to relieve or gratify certain sensations, desires, etc.; and whenever the same state of mind is induced, however feebly, there is a tendency through the force of habit and association for the same movements to be performed, though they may not then be of the least use.[107]

The idea here is that some instances of expressive behaviour originally served *non-expressive* functions—such as relieving itches, avoiding pains, or gratifying basic desires—but eventually became expressive when the mental states that caused the original behaviour were later 'associated' with similar mental states. For example, when children wish to distance themselves from a disagreeable object they often wilfully shove it away from themselves. This vigorous extension of an arm is directly serviceable in gratifying the child's basic desire to be rid of the unwanted object. As Darwin observes though, we tend to automatically perform the same shoving action even when it is clearly of no use: "A man or a child in a passion, if he tells anyone in a loud voice to be gone,

---

[106] Ibid., p. 12.
[107] Ibid., p. 34.

generally moves his arm as if to push him away, although the offender may not be standing near."[108] Here the automatic shoving gesture fails to serve its original function of physically distancing oneself from the unwanted object. Neither, apparently, does it serve any novel function; it serves no purpose beyond expressing our desire. We now act in this way, though, because our current state of mind is similar to the state of mind in which, as children, our wilful shoving was practically effective. Similar states of mind thus initiate similar forms of behaviour even though certain instances of the behaviour fail to perform their original function, or serve no function whatsoever.[109]

At this point three features of Darwin's account of his first principle become immediately apparent. First, notice in the above example that the original serviceable act—shoving the object away—was at first *consciously* performed. This is true, Darwin argues, of all expressive behaviours explicable under the first principle.[110] Moreover, they were consciously performed for some practical end *other than expression*. Later, as these originally conscious behaviours became *habits* through constant repetition, they came to be performed automatically and unconsciously, *only then* serving as expressions.[111] As presented by Darwin, however, this picture contains two fundamental problems. First, in elaborating his first principle Darwin often relied on the now

---

[108] Ibid., p. 67.

[109] Darwin repeatedly emphasizes that expressive behaviours were often 'of no use.' Any serious consideration of expression's value as a form of *communication* is conspicuously absent. Richard Burkhardt suggests that, in part, Darwin chose not to emphasize expression's communicative value because he was concerned to allow that not all characteristics were necessarily adaptive. Darwin had made this point in earlier works, but in writing *The Expression of the Emotions* he seemed particularly sensitive to the issue (Burkhardt 1985, pp. 357-59). I consider two other possible explanations below.

[110] "All [expressions] included under our first principle were at first voluntarily performed for a definite object" (ibid., p. 349).

[111] Darwin argued that habits further shaded indistinguishably into *reflexes*, the difference between the two depending upon the degree to which higher brain centres became involved in the processing of the stimulus. See pp. 41-43.

discredited Lamarckian view that behaviour acquired by our ancestors as habits could be transmitted to offspring in the form of *instincts*, i.e., as innate unconscious behaviour.[112] In the end, however, this does not pose particular problems for Darwin—the existence of variation in reflex renders expression subject to natural selection.[113]

A second and apparently more significant problem in Darwin's first principle is that it seems to exclude emotions from playing *any* causal role in the original production of behaviours expressive of emotion. Emotions, in short, are rendered explanatorily redundant. John Dewey formulated the criticism succinctly:

> ...the principle of explanation *actually* used, whatever the form of words employed, is that of survival...of acts originally useful not *qua* expressing emotion, but *qua* acts—as serving life....*The reference to emotion in explaining the* [behaviour] *is wholly irrelevant; the attitude of emotion is explained positively by reference to useful movements.*[114]

Dewey's point is a simple one. According to Darwin, emotions *qua* mental states nowhere figure in the production of the originally serviceable behaviours that have only lately become expressive and hence non-serviceable. Emotions thus do not serve to explain the *origins* of those behaviours. It is only at a later stage, once rendered habitual and inherited as instinct, that these original behaviours come to be connected with particular emotions *as expressions* of those emotions. Emotions only later explain the appearance of particular behaviours, and this because our current emotions qua mental

---

[112] For an example of Darwin's Lamarckism, see his explanation of blinking in the startle reaction, p. 45. For Darwin's identification of instinct with innateness, see pp. 22 and 124.
[113] Ibid., p. 47; see also p.36. While ultimately unproblematic, Darwin's reliance upon Lamarckism is still worth noting as it emphasizes the importance of heritability in Darwin's arguments against Bell. Darwin must argue specifically for the heritability of expression, otherwise the universality of expression across races and species would not serve as evidence for a common progenitor.
[114] Dewey 1894, pp. 154-555; italics in the original.

states bear some similarity to the non-emotional mental states that originally produced the behaviour now under consideration.

While later commentators, Dewey included, have seized on this apparent explanatory redundance as a significant flaw in Darwin's account, such criticism is misguided. To rightly point out that emotions qua mental states are explanatorily redundant in Darwin's account of the origins of particular expressive behaviours is not to show that Darwin was wrong in his explanations. Darwin's account must be evaluated on its own terms. Simply put, was he correct in asserting that behaviour now read as expressive was originally selected for its non-expressive functions? Critics of Darwin's narrowness, however, generally don't fault him in the details of his account.[115] Sue Campbell, for example, argues instead that Darwin's commitment to a narrow class of involuntary behaviours characteristic of the most basic emotions precludes him from assigning emotions any explanatory role in more complex, intentional expressive behaviours.[116] But clearly there is no *prima facie* reason for supposing this to be so.

Still, Dewey and Campbell have raised an important point. Emotions qua mental states are inert in Darwin's explanation of the origins of expressive behaviour. They are not similarly inert, though, in explaining 'modern' instances of expressive behaviour. For Darwin, as the example below shows, emotional states such as grief and anxiety do in fact *now* cause us to furrow our brows, even though such behaviour originally served a protective as opposed to expressive function. The value of Dewey's critique is that it shifts our focus to the process whereby emotions *became* involved in the production of

[115] In fact, many of Darwin's explanations are now widely accepted. For a good sense of how Darwin has stood up in this regard, see Ekman's comments throughout the text, especially pp. 54, 72, 75, and 79.

behaviour. When highlighted in this way, Darwin's answer—which relies on an unanalysed notion of 'association of similar mental states'—is seen as radically incomplete. Having noted this fundamental problem, however, it is possible to see ways in which Darwin might work out a solution. For example, in his account of the origins of blushing—which I will discuss below—he argues that the mental states that first caused blushing are similar to the modern, specifically emotional causal states in that both involve a form of self-attention. In the earlier states this attention was directed toward one's physical appearance; in the later states the self-attention is more complex and involves a moral dimension. Darwin's appeal to similarity of mental states thus need not remain mysterious. But more about this later.

Turning now to the final point I want to make about the principle of serviceable associated habit, notice how Darwin's claim for the non-expressive origins of expressive behaviour speaks directly against Bell's creationist doctrine that God had endowed humans with *uniquely expressive* musculature. For Bell and other natural theologians of Darwin's time the human capacity for expression was a marker of humanity's discontinuity with nature. If Darwin could prove that our expressive behaviour actually had its origins in behaviour that first served more prosaic functions—especially functions that we clearly shared with the 'lower' species—he would be one step closer to disproving natural theological claims about human distinctness. A great deal thus rested on whether Darwin could prove the basic claim encased in the principle of serviceable associated habit. The best way to prove this claim was to tell a plausible story of how expressive behaviour could arise from non-expressive behaviour, and the more 'purely

---

[116] Campbell 1997, p.22. Campbell nowhere challenges Darwin's explanations.

expressive' the original behaviour being explained apparently was, the better. Darwin's case against Bell would thus be strongest were he able to explain the origins of behaviour that seemed to never have had—and never could have had—any but a purely expressive function. Darwin thus went to his greatest lengths in explaining expressive behaviours that seemed particularly trivial or useless.

Consider, for example, Darwin's ingenious answer to a question that had long troubled him: "During several years no expression seemed to me so utterly perplexing as this one which we are here considering. Why should grief or anxiety cause the central fasciae alone of the frontal muscle together with those round the eyes, to contract?"[117] The contraction of these muscles produces an expression that Darwin argues is characteristic of grief—an oblique slanting of the eyebrows and a furrowed central brow. He notes: "Here we seem to have a complex movement for the sole purpose of expressing grief."[118] It is thus a perfect example for Darwin's case against Bell, since it is hard to see what non-expressive function this particular bit of peculiar behaviour could ever have served.

Darwin's answer to the perplexing question proceeds from the apparently unrelated observation that when one tilts their head upward to view a strongly illuminated surface, the orbicular, corrugator, and pyrimidal muscles are automatically contracted so as to prevent damage to the eyes. Darwin further noticed that when subjects struggled against these natural protective contractions, in an attempt to maintain eye contact, this struggle brought into play the central fascia muscles such that the struggle of these

---

[117] *The Expression of the Emotions*, pp. 186-7
[118] Ibid., p. 187.

antagonistic muscles against the natural contractions around the eye produced the expression characteristic of grief—an oblique slanting of the eyebrows and a furrowed central brow. Darwin then noted that crying or screaming children contract their orbicular, corrugator, and pyrimidal muscles in exactly the same way as subjects looking upward into a light. The reason for this, Darwin had previously argued, is that contracting the muscles in this fashion prevents the eyes from damage by over-engorgement.[119] These similarities led Darwin to expect that when children attempted to stop crying or screaming, the same antagonistic muscles should be brought into play, thereby producing grief's characteristic expression. Observation confirmed his expectation.[120] This observation, Darwin writes, supplied "the key to the problem":

> We have all of us, as infants, repeatedly contracted our orbicular, corrugator, and pyramidal muscles, in order to protect our eyes whilst screaming...and though with advancing years we easily prevent, when feeling distressed, the utterance of screams, we cannot from long habit always prevent a slight contraction of the above named muscles;...their contraction can be checked only by the antagonistic contraction of the central fasciae of the frontal muscle. The result which necessarily follows...is the oblique drawing up of the eyebrows, the puckering of the inner ends, and the formation of rectangular furrows on the middle of the forehead.[121]

Grief's characteristic expression thus springs from behaviour that was in the first place essentially *protective*. When infants cry or scream, contractions of the muscles around the eyes serve only to protect the eye from damage by engorgement. Infants, however,

---

[119] Darwin credited Bell with first proving this function (p. 7). Darwin also himself argued this point at length in support of Bell (pp.159-164). Ekman suggests the care Darwin took in establishing this point stems from the central role it plays in his argument about the origins of grief's characteristic expression (p. 161).

[120] Darwin enlisted Henrietta Huxley, Thomas' wife, in these observations, and indicated to her what she might expect to see. In a letter to Thomas Huxley, Darwin wrote: "Ask [Mrs Huxley] to look out when one of her children is struggling and just going to burst out crying....A dear young lady near here plagued a very young child for my sake, till it cried, and saw the eyebrows for a second or two beautifully oblique, just before the torrent of tears began" (quoted in Browne 1985, p. 307).

generally only cry or scream when distressed, so over time the mental state of distress comes to be associated with crying, and hence with the associated protective contractions. As adults, though, we are able to check our tears through either habit or force of will. Exerting control in this way activates muscles antagonistic to those that are protectively contracted. In this way we produce grief's characteristic expression.[122] The essential point here, for Darwin, is that in its original form this 'expressive' behaviour served only a *protective* and hence non-expressive function.

Darwin's other principles similarly argue for expressive behaviour's non-expressive foundations. His second principle— the *principle of antithesis*—is essentially an extension of the principle of serviceable associated habit: "When a directly opposite state of mind is induced, there is a strong and involuntary tendency to the performance of movements of a directly opposite nature, though these are of no use; and such movements are in some cases highly expressive."[123] Darwin's draws his examples here, which are mainly of animals' expressive behaviours, from exquisite observation. He notes, for example, that when dogs are in a hostile frame of mind—as when they approach strange dogs or humans—they typically effect a distinct behavioural set. They walk stiffly with an upraised head; their tail is held rigidly erect; hair along their neck and back bristles; their ears are pricked and turned forward; their eyes are set in a fixed stare; and their canine teeth are bared. If, however, the dog suddenly recognises the stranger as his master, this behavioural set is immediately 'reversed.' Darwin describes the change:

---

[121] *The Expressions of the Emotions*, pp. 189-90.
[122] It is perhaps better to say that in this way humanity's animal progenitors produced grief's characteristic expression. Modern humans produce this expression because we have inherited our progenitors' habits in the form of instincts.
[123] Ibid., p. 34.

"Instead of walking upright, the body sinks downwards, and is thrown into flexuous movements; his tail...is lowered and wagged...his hair instantly becomes smooth; his ears are depressed and drawn backwards...and his lips hang loosely."[124] A similar 'reversal' of behaviour is observed when cats move between aggressive and affectionate frames of mind.[125]

For Darwin these friendly 'reversed' forms of aggressive behaviour are inexplicable in terms of his first principle of serviceable associated habit. While the aggressive behaviours are so explicable—because they were originally, and still are, of some direct or indirect service to the animal in times of danger—the movements "so clearly expressive of affection" seem never to have been of any service whatsoever.[126] This leads Darwin to the conclusion that such behaviour is explicable solely in terms of its being "in complete opposition or antithesis" to the behaviour exhibited when the subject is in an 'opposite or antithetical' frame of mind.[127] To make such a claim, though, is merely to argue that certain behaviours are unified by a single principle. It is not to explain *why* that principle holds when it does

As a possible first candidate for explaining why 'solely antithetical' behaviours exist, Darwin considers their potential as forms of communication. Some gestures, being "manifestly of an opposite nature to those by which certain feelings are already expressed" might once have been consciously performed with the intention of

---

[124] Ibid., p. 56.
[125] Ibid., pp. 59-60.
[126] Ibid., p. 56. For Darwin's account of the adaptive value of canine expressions of hostility see pp. 116-18.
[127] Ibid., p. 56.

communicating opposite states of mind.[128] Our ancestors might have chosen to behave in a certain way because that behaviour, being 'manifestly opposite' to our functional aggressive behaviour, communicated to others that no harm was meant. These conscious behaviours, through some Lamarckian mechanism of inheritance, could then have become the innate and automatic 'reverse' behaviours we exhibit today. Darwin insists, however, that in cases where the original 'reverse' behaviour might have served a communicative function, that behaviour must have been performed *consciously*. For example, in considering the communicative value antithetical behaviours do in fact possess, Darwin points out that "the fact of the gestures being now innate, would be no valid objection to the belief that they were at first intentional."[129] Relatedly, Darwin considers the communicative explanation for the principle of antithesis only to dismiss it on the grounds that it is highly unlikely that the first instances of opposite behaviour were originally consciously performed. It would be incredible, Darwin argues, to believe that friendly dogs originally exhibited the 'reverse aggressive' behaviour they do because they were aware that that behaviour would be read as indicating friendly intent.[130] Communication, for Darwin, requires conscious intent, and so without the conscious intent to communicate on the part of our ancestors the communicative value of antithetical behaviours could not have been instrumental in their evolution.

Having so eliminated communicative value as an evolutionary factor in this case, Darwin is reduced to claiming that the principle of antithesis is simply a irreducible principle of nature. He concludes his discussion of the principle: "When actions of one

---

[128] Ibid., p. 63.
[129] Ibid., p. 63.

kind have become firmly associated with any sensation or emotion, it *appears natural* that actions of a directly opposite kind, though of no use, should be unconsciously performed through habit and association, under the influence of a directly opposite sensation or emotion."[131] Darwin does not want, however, to completely discount the possibility that communicative value might have played some role in the evolution of antithetical expressive behaviours. He allows that our irreducible natural tendency to perform such behaviours might have been strengthened by their communicative value: "If indeed [antithetical behaviours] are serviceable to man or to any other animal, in aid of inarticulate cries or language, they will likewise be voluntarily employed, and the habit will thus be strengthened."[132]

Darwin's ambiguous positioning of the role of expression's communicative value bears comment. Clearly, Darwin saw that the expressive behaviours he was considering were significant forms of communication. In his summary discussion of the importance of expression he wrote:

> The movements of expression in the face and body, whatever their origin may have been, are in themselves of much importance for our welfare. *They serve as the first means of communication between the mother and her infant*; she smiles approval, and thus encourages her child on the right path, or frowns disapproval. We readily perceive sympathy in others by their expression; our sufferings are thus mitigated and our pleasures increased; and mutual good feeling is thus strengthened. The movements of expression give vividness and energy to our spoken words. *They reveal the thoughts and intentions of others more truly than do words, which may be falsified.*[133]

---

[130] Ibid., p. 66.
[131] Ibid., p. 67; my italics.
[132] Ibid., p. 67.
[133] Ibid., p. 359; my italics.

Expression forms the basis of infant communication; it serves adults in establishing and maintaining communion with others; it even trumps language in perspicuity.[134] Why then was Darwin so reluctant to consider selection for communicative value as a fourth principle? Two answers suggest themselves.

First, de-emphasising expression's worth as a form of communication was likely a strategic rhetorical move. As previously noted, Bell and other natural theologians—who all were the main target of Darwin's work on expression—had argued that the capacity for expression had been uniquely given to humans precisely *for the purpose* of communication. Bell made the point clearly:

> ...in man there seems to be a special apparatus, for the purpose of enabling him to communicate with his fellow creatures, by that natural language which is read in the changes of his countenance. There exist in his face, not only all those parts which by their action produce expression in the several classes of quadrupeds, but there is added a peculiar set of muscles to which no other office can be assigned than to serve for expression.[135]

As exemplified by Bell the natural theological position here consisted of three basic claims: humans possess a unique expressive musculature; the purpose of this musculature is communication; the uniqueness of this capacity is proof of humanity's discontinuity. To place too much emphasis on expression as a form of communication would then have aligned Darwin too closely with the position against which he was arguing. De-emphasising expression's communicative value thus served to distinguish him from the

---

[134] Darwin suggests elsewhere that some behaviours serve solely as signals of danger (p. 130) and as signals of sexual availability (p. 98).
[135] Quoted in Burkhardt 1985, p. 358.

opposition.[136] I also want to suggest, though, a second and more instructive reason for Darwin's lack of emphasis on the communicative value of expression.

For behaviours to be selected for their value in communicating inner feelings and intentions to act, two general conditions must hold. First, particular forms of putatively expressive behaviour must, on the whole, be *stable* and *reliable* indicators of the feelings or intended actions they are read to be expressive of. If the submissive behaviours of a friendly dog are to be selected for their value in indicating friendly intent, then those behaviours must (1) be replicated consistently enough so as to be identifiable by others (stability), and (2), they must, *on the whole*, be followed by friendly actions (reliability).[137] Were either condition not to hold then a particular form of behaviour would not be able to function as a form of communication. Explaining how such stable and reliable connections could have arisen, though, was not a particular problem for Darwin— the principles he proposes do exactly that. The second condition that must have held, though, was more problematic.

To have been selected for their communicative value, expressive behaviours *qua* symbols of intended action must also elicit relatively stable and reliable *reactions* from the *consumers* of those symbols; expressions must be *on the whole* consistently

---

[136] Richard Burkhardt notes that this reactionary strategy led to a significant missed opportunity: "In constructing his argument against the idea that special structures in man had been designed by the creator for the purpose of non-verbal communication, Darwin appears to have overreacted, thereby leaving himself ill-disposed to develop an idea that would later be advanced by the ethologists of the twentieth century— the idea that certain expressive actions, whatever their primary origin, had been developed over time by natural selection" (ibid., p. 360).

[137] Of course, there can be significant survival value in misleading others through *feigned* intent. Feigning, however, can only work against a larger 'history of truth-telling', i.e., of reliable indication. This is the lesson of the 'Boy Who Cried Wolf.' This history of truth-telling, however, need not be the history of a particular individual. Some species—e.g., certain varieties of sea anenomes—make their living off consistently *misrepresenting* themselves as having friendly intentions. They succeed in this, however, only by imitating the behaviour of other species that *is* a reliable indicator of friendly intent.

interpreted. If consumers of symbolic expressive behaviour failed to respond consistently to instances of symbolic behaviour, then that behaviour would be of little advantage to the expressing organism.[138] If a dog's expressions of friendly intent were *on the whole* misinterpreted by other dogs as signalling hostility, and subsequently acted on as such, then those expressions would have little or no communicative value. Similarly, if a piece of expressive behaviour elicited wildly variant responses from the 'consumers' of the behaviour, it would be of little use as a piece of communication. A consistent reaction to a consistently produced symbol, though, would likely have led to habitual responses on the part of the symbol's interpreter, habits that in turn would have been rendered instinctual. The capacity for expression would then likely have evolved lockstep with the related capacity to properly interpret expression. As expressive behaviours were rendered innate and instinctive over time, so too would have the ability to recognise the meanings of those expressions. Proving this to be the case, though, posed a problem for Darwin.

Darwin raised the question of an instinctive capacity for such recognition only in his last chapter, but his few remarks there are illuminating. Darwin thought it likely that recognition and expression had in fact evolved simultaneously: "As most of the movements of expression must have been gradually acquired, afterwards becoming instinctive, there seems to be some degree of *a priori* probability that their recognition would likewise have become instinctive."[139] The motivation behind this claim was Darwin's observation that animals and children seemed to instinctively recognise the significance of basic expressions. Monkeys and dogs, for example, seemed to understand

---

[138] Millikan 1984, p. 30.
[139] *The Expression of the Emotions*, p. 353.

the significance of smiles, laughs, and threatening tones of voice. Darwin allowed, however, that many of these cases might be explicable as instances of learning. To establish an instinctive capacity for recognition Darwin thus turned to the study of his first-born infant. Observing a child's unlearned responses, he reasoned, would provide the strongest proof for the existence of an instinctive capacity for recognition. His studies were less than conclusive though: "It is extremely difficult to prove that our children instinctively recognise any expression."[140] In the end, however, Darwin's observation of his child's sympathetic reaction to a nurses pretend tears—the infant depressed the corners of his mouth—led him to only a tentative conclusion: "It seems to me that an innate feeling must have told him that the pretended crying of his nurse expressed grief: and this, through the instinct of sympathy, excited grief in him."[141]

I would suggest, then, that Darwin's reluctance to consider selection for communicative value as a fourth principle stems in part from his difficulties in determining the extent to which the capacity to recognise the significance of emotional expressions was instinctive. He saw, I would argue, that expression's having been selected *for* its communicative value would likely—perhaps even necessarily—have entailed a lockstep selection of capacities for correct interpretation. To make the point in a slightly different way, for a piece of behaviour to have originally been selected for its expressive power there would have had to have been a stability and reliability in the reactions of that behaviour's consumers, the existence of which would have been difficult or even impossible to prove. In the end, Darwin simply found it easier to argue for

---

[140] Ibid., p. 353.
[141] Ibid., p. 354.

originally non-expressive functions such as protection and avoidance of danger. I would suggest, then, that taken together these difficulties presented Darwin with good reason for side-stepping the issue of communication. And when joined with the other reasons noted above, the conspicuous absence of a thorough consideration of expression's communicative value is understandable. Darwin was thus content to search for the adaptive value of expression's originating behaviours almost exclusively along other, non-communicative dimensions.

This brings us to the final principle of expression proposed by Darwin. Like the first two principles, Darwin's third—the *principle of direct action of the nervous system*—posits a non-expressive foundation for expressive behaviour. The claim here is that some expressive behaviour is purely a function of the mechanics of human neurophysiology:

> When the sensorium is strongly excited, nerve force is generated in excess, and is transmitted in certain definite directions, depending on the connection of the nerve-cells, and partly on habit: or the supply of nerve force may, as it appears, be interrupted. Effects are thus produced which we *recognise as expressive.*[142]

Key examples include the trembling of muscles under fear and extreme pleasure; changes in digestive function and in the activity of glands under the influence of strong emotions; and fluctuations in heart rate.[143] These phenomena, Darwin argues, are all the result of the essentially 'hydraulic' nature of the nervous system. "Nerve force," like fluid, travels along pre-established nervous pathways and when generated in excess—as in times of high emotion—it 'spills over' into the neural equivalents of storm drains and flood plains. As these phenomena are purely the result of the physical constitution of the nervous

---

[142] Ibid., p. 34.

system, Darwin notes, it follows that the behaviours so produced have never been under conscious control, unlike the behaviours explicable by the first and second principles.[144] These entirely unconscious behaviours, however, can be indirectly influenced by the will through the physical effects of habit; nervous excess travels most easily along neural pathways which have been strengthened by habit.[145] Darwin makes use of this combination of principles in explaining blushing, an expression that played a central role in his argument against Bell.

Blushing presented a unique challenge to Darwin. It is, he noted, "the most peculiar and the most human of all expressions. Monkey's redden from passion, but it would require an overwhelming amount of evidence to make us believe that any animal could blush."[146] Blushing thus appears to mark a true discontinuity between humans and other species. Bell and other natural theologians of the time, for example, explained blushing as a special creation intended for the purpose of expressing specifically *moral* feelings.[147] Blushing was thus unique because man alone was capable of moral feeling.

Not surprisingly, Darwin's explanation of the origins of blushing eschewed appeals to a uniquely human moral sense. His account instead rested upon the morally neutral notion of self-attention. Drawing on medical authority, Darwin first correctly argued that concentrated self-attention focused upon any particular body part tends to

---

[143] Note that, strictly speaking, internal physiological changes cannot be expressive.
[144] Ibid., p. 69.
[145] Ibid., p. 74.
[146] Ibid., p. 310.
[147] Darwin quotes Thomas Burgess, who argued that the blush had been designed by God "in order that the soul might have sovereign power of displaying in the cheeks the various internal emotions of the moral feelings" (ibid., p. 335). This particular claim has found sinister expression in the modern racist doctrine of "Blood in the Face," which holds that because races with dark skin can't blush they are incapable of experiencing shame and hence are morally inferior. Not surprisingly, in addition to being morally

interfere with the muscle tone of arteries and surface capillaries in that part, thereby causing those vessels to relax and fill with arterial blood.[148] This sudden flooding appears on the skin's surface as blushing's reddish bloom. Precisely *why* self-attention should interfere in this way with normal bodily functions ultimately remained mysterious for Darwin. He canvassed several possible solutions but was content to leave the question unanswered, concluding that it is most likely just a result of our neurophysiological constitutions.[149]

Darwin did not need, however, to explain the physical effects of self-attention; he only needed to prove their existence. Having done so he could then argue that the original state of mind that first induced blushing in our ancestors was a simple "self-attention directed to *personal appearance*, and not to moral conduct."[150] His arguments for this claim are varied, but all essentially work toward establishing the ubiquity of a deeply embedded concern for personal appearance that is distinct from, and evolutionarily prior to, a uniquely *moral* sensibility.[151] Having shown that simple self-attention to appearance is primary, Darwin then goes on to argue that over time blushing came to be induced by more complex states of "self-attention in relation to moral

---

repugnant, the factual claim at the heart of the doctrine is wrong. Darwin himself observed through numerous informants the appearance of blushing in a range of dark-skinned races (pp. 315-319).

[148] Ibid., p. 336.

[149] Ibid., pp. 336-42.

[150] Ibid., p. 324.

[151] Darwin cites a long list of observations: "It is notorious that nothing makes a shy person blush so much as any remark, however slight, on his personal appearance. One cannot notice even the dress of a woman much given to blushing, without causing her face to crimson. It is sufficient to stare hard at some persons to make them, as Coleridge remarks, blush—'account for that he who can'. With...two albinos observed by Dr. Burgess, 'the slightest attempt to examine their peculiarities invariably caused them to blush deeply. Women are much more sensitive about their personal appearance than men are, especially elderly women in comparison with elderly men, and they blush much more freely....It is plain to every one that young men and women are highly sensitive to the opinion of each other with reference to their personal

conduct."[152] He provides a survey of these evolutionarily newer causes of blushing—they include shyness, guilt, breaches of etiquette, and modesty—but argues that *all contain the original and essential element of self-attention.*[153] Here, for example, is Darwin's explanation of why guilt leads to blushing:

> With respect to blushing from strictly moral causes, we meet with the same fundamental principle as before, namely regard for the opinion of others. It is not the conscience which raises a blush, for a man may sincerely regret some slight fault committed in solitude, or he may suffer the deepest remorse for an undetected crime, but he will not blush. 'I blush,' says Dr. Burgess, 'in the presence of my accusers'. It is not the sense of guilt, but the thought that others think or know us to be guilty which crimsons the face. A man may feel thoroughly ashamed at having told a small falsehood, without blushing; but if he even suspects that he is detected he will instantly blush, especially if detected by one whom he reveres.[154]

The first mental states that caused blushing are thus related to the newer and more complex eliciting states in virtue of their being similar *patterns of focus and attention upon the self.* Darwin sums up his account:

> I conclude that blushing—whether due to shyness—to shame for a real crime—to shame from a breach of laws of etiquette—to modesty from humility—to modesty from an indelicacy—depends in all cases on the same principle; this principle being a sensitive regard for the opinion, more particularly for the depreciation of others, primarily in relation to our personal appearance, especially of our faces; and secondarily, through the force of association and habit, in relation to the opinion of others on our conduct.[155]

Thus on Darwin's conception the emergence of blushing in humans is not proof of a uniquely human, God-given moral sense, but is rather proof only of a radically increased capacity and propensity for richer and more complex forms of self-awareness.

---

appearance; and they blush incomparably more in the presence of the opposite sex than in that of their own" (p.325).

[152] Ibid., p. 324.

[153] "*The nature of the mental states which induce blushing.* These consist of shyness, shame, and modesty; the essential element in all being self-attention" (p.324).

[154] Ibid., p. 331.

This is a good point at which to leave Darwin, for in his account of blushing we see nicely summed up the fundamental goal and strategy of his entire study of emotional expression. *Contra* the tradition of natural theology, which saw human emotional expression as evidence of a radical discontinuity with nature, Darwin argued that human expression instead reveals our deep continuity with the lower animals. He supported this claim by showing that the best explanation of a range of human expressive behaviours is given by locating the origin of those behaviours in the evolutionarily ancient, *non-expressive* behaviours shared by humans and other species.

With this account of blushing we also see how Darwin might begin to fill in the most serious lacuna in his theory, namely, the lack of any real explanation of how emotions *qua* mental states eventually came to be involved in the production of these behaviours. As noted above, Darwin typically answers this question by relying on an unanalysed appeal to *similarity of eliciting mental states*. In the case of blushing, however, we see Darwin edging toward a more substantial account. Here, the relevant similarity between the early and later elicitors of blushing is a *similarity of focus and attention*.

Conclusion

I want to close this chapter by suggesting that the history offered here shows an important progression of thought, one that is mirrored in the account of the modern theories that I turn to now, and one that I hope to extend. To begin, we see in Plato—at

---

[155] Ibid., p. 334.

least in his tropes and metaphors—an influential but unsatisfactory division between emotion and reason. They are portrayed as distinct faculties that often come into conflict. Following Plato, Aristotle rejects this division by showing how closely emotion is tied to judgement. On Aristotle's view, in fact, our emotions are partially *constituted* by certain judgements, and much of his account of emotion is spent constructing a 'formal' model that details the relations between particular emotions and particular classes of these judgements. Aristotle is thus one of the first 'cognitive theorists of emotion,' a tradition which I will examine in more detail in the next chapter. Missing from Aristotle's otherwise thorough account, however, is any serious discussion of the *function* of the emotions. Turning to Descartes, however, we do see a serious and sustained attempt to explain the function of emotion. In Descartes' account our emotions serve to promote a beneficial alignment of mind and body in response to highly significant situations. This account of function, in turn, allows Descartes to answer an important question left unasked by Aristotle—it allows him to explain *why* certain classes of judgements, and not others, are tied to particular emotions. Of course, Descartes cannot ultimately explain why we have emotions at all, nor can he tell us much about *how* our emotions perform the functions that they do. The task of answering these deeper questions instead falls to Darwin's close empirical investigations and the theory of natural selection.

This is a special way of being afraid
No trick dispels. Religion used to try,
That vast moth-eaten musical brocade
Created to pretend we never die,
And specious stuff that says *No rational being*
Can fear a thing it will not feel, not seeing
That this is what we fear – no sight, no sound,
No touch or taste or smell, nothing to think with,
Nothing to love or link with,
The anaesthetic from which none come round.

- from *Aubade*, by Phillip Larkin[1]

Curing The Fear Of Death

The morbidly inclined among us will be relieved to discover that counter to Larkin's poetic claim the fear of death has in fact been dispelled. And not once, but twice, by tricks philosophical and pharmacological.

The latter cure is the (alas) fictional centrepiece of Don DeLillo's brilliant *White Noise*, a novel that traces the fate of a Middle American family whose existence is forever disrupted when an industrial accident unleashes a lethal "airborne toxic event" over their hometown. In the following passage Jack Gladney, a professor of "Hitler studies," has just discovered his wife Babette's infidelity. She has admitted to sleeping with a representative of Gray Research, a pharmaceutical company, so as to be able to obtain advance samples of their newest drug, Dylar. Babette explains to her bewildered husband:

> "They isolated the fear-of-death part of the brain. Dylar speeds relief to that sector."

[1] Thanks to Ronnie de Sousa for bringing this poem to my attention.

> "Incredible."
>
> "It's not just a powerful tranquilizer. The drug specifically interacts with neurotransmitters in the brain that are related to the fear of death. Every emotion or sensation has its own neurotransmitters. Mr. Gray found fear of death and then went to work on finding the chemicals that would induce the brain to make its own inhibitors."
>
> "Amazing and frightening."
>
> "Everything that goes on in your whole life is a result of molecules rushing around somewhere in your brain.'"[2]

Dylar is thus a drug of the purest function, zeroing precisely in on the 'fear-of-death part of the brain,' working its magic *only there*. Babette takes the drug and retains all of her normal cognitive functions. With the exception of her obviated fear she thinks and feels just as before. Her beliefs, desires, and attitudes about death remain wholly intact, as do all her other beliefs, desires, and attitudes. Her other fears similarly remain unchanged. And yet she no longer fears death.

Could such a drug exist? Could we truly retain all of our beliefs, attitudes, wishes and desires—in general, all of our cognitive states—concerning death, and yet cease to fear it by ingesting a dose of Dylar?[3] The answer is left as an exercise to the reader, but it is a question worth keeping in mind throughout the following sections.

The philosophical cure is found in Robert Gordon's *The Structure of Emotions*, and while it is never explicitly prescribed there it is easily derived from the theory of emotion developed within. In this work Gordon argues that emotions fall neatly into two kinds: *factive* and *epistemic*. A particular emotion's classification depends upon essentially epistemic facts, i.e., on how it relates to beliefs, knowledge, and facts about the world. Factive emotions—anger, shame and sadness are the main factive emotions—

---

[2] DeLillo 1986, p. 200.

[3] For some relevant thoughts on the "a priori limits of psychopharmacology," see Gordon 1987, pp. 49-52.

are those which *require* belief and/or knowledge.[4] Consider, for example, Gordon's 'Belief Condition' on anger:

> **BC:** If S is angry about the fact that p, then S's believing that p is sufficient for S to be angry, given some existing conditions that are not themselves sufficient for S to be angry.[5]

The condition is intuitively appealing. It seems only right that if I am angry that my dog ate my slippers then **I** must believe that the dog ate my slippers. The claim "I'm angry that the dog ate my slippers but I don't believe he ate them' is *prima facie* inconsistent.

Conversely, epistemic emotions – fear, hope and worry are the sole epistemic emotions – are those which *preclude* knowledge. For any epistemic emotion " 'S emotes (e.g., is afraid, is hopeful, is worried) that *p*' is true only if S is not certain that *p* (and therefore cannot said to know that *p*)."[6] Again, the condition is intuitively appealing. It would sound decidedly odd were someone to claim that they hoped they would find their lost dog even as they expressed their happiness over actually having found it. Similarly, it seems in some sense 'impossible' to claim that I am worried that the plane will crash even as I admit that I know that it has just landed safely and rolled to a complete stop. In both cases an agent's knowledge that certain facts obtain seem to preclude that agent's experiencing certain emotions about those facts.

Of course, in all cases, an emotion requires more than simply an agent's being in a certain epistemic state. Gordon's condition **BC** on anger, for example, only claims that belief is one component of anger's necessary preconditions. These added conditions,

---

[4] For any factive emotion " 'S emotes (e.g., is amazed, is angry, is delighted) that *p*' is true if and only if it is true that *p* and, further, that S knows that *p*" (Gordon 1987, p. 43).
[5] Ibid., p. 48.
[6] Ibid., p. 43.

Gordon argues, take the general form of negative or positive *attitudes* toward the object of one's emotion. These attitudes, Gordon further argues, are for various reasons best construed as being *wishes*.[7] A fuller account of any emotion thus requires, in addition to a description of its necessary epistemic conditions, a specification of that emotion's necessary attitude. A complete specification of the preconditions for fear runs as follows:

> **R**: First, if S fears or is afraid or terrified that *p*, then S *cares* whether or not *p*: More specifically, S wishes it not to be the case that *p* ('wishes that not-*p*,' for short). And second, if S fears (is afraid) that *p*, then *S is neither certain that* p *nor certain that not*-p.[8]

Given **R**, consider now a subject S contemplating his own death. According to Gordon, S fears death—or more specifically 'S fears that he will die'—if and only if:

> (1) S wishes it not to be the case that he will die.
> (2) S is neither certain that he will die, nor certain that he won't.

Suppose now that (1) is true, as it most often is: S wishes not to die. The second requirement for S to fear death, however, is problematic. Assuming S to be a well informed rational agent he likely understands his death to be an inescapable certainty. The second condition thus fails to hold and so by **R** and *modus tollens* whatever state of mind S might be in regarding his death it cannot be the state of fearing that he will die. The certainty of death precludes the fear that it will occur.

---

[7] Ibid., p. 30. Gordon rejects the standard view that the relevant additional attitudes of emotions are 'wants' or 'desires' for the following reason. When S is happy about the fact that *p*, then on Gordon's view S must know that *p*. But 'knows that *p*' implies that *p* is already the case. And if *p* is already the case, Gordon claims, then this "leaves S with no possibility of instrumental or aversive action. In such a case the notions of wanting and desiring are usually thought inapplicable" (ibid.). For example, if it is true that I know that the plane has landed safely, then it must be true that the plane has indeed landed safely. If this is so, however, it seems a deviant usage of 'want' to claim that I might now *want* the plane to *not* have landed safely as there is no action I can now perform to bring this about. Wishing, Gordon argues, is thus better suited as a description of emotion's required positive or negative attitude since it has a high degree of logical transparency. I can, for example, wish things to have been different than they actually are; I can wish that the plane crashed even while I know it landed safely.

[8] Ibid., p. 68.

This strikingly odd conclusion undoubtedly flows from the fact that the fear of death is here being construed *propositionally*. S is being denied the fear '*that* he will die,' not the fear '*of* death.' Again, Gordon's general point underlying this reading sounds intuitively plausible in certain examples. Just as in the cases of hope and worry, it seems 'impossible' in some sense for me to fear *that* the plane I'm flying in will crash even as the plane's rolling to a stop convinces me that it hasn't crashed. Were I to make this claim I would undoubtedly be challenged. But why is this so? In exactly what sense is this state of mind 'impossible'? Here, 'impossible' is ambiguous between at least two importantly different senses: the logical and the psychological. On the weakest reading, we might understand the outcome of Gordon's stricture **R** as pointing to the logical contradiction of attributing to an agent the incompatible mental states of fearing his death when he knows it to be certain. But Gordon's point is not in the first place concerned with attribution. He presents **R** as a condition for the existence of the mental state 'fearing that *x*,' not as a condition for *attributing* that state, and while there are obvious ties the issues are distinct. In a logical sense, then, the 'impossibility' of fearing death on this reading might seem analogous to the *irrationality* of believing a contradiction. If this is so, however, the philosophical cure is a weak remedy, for it is surely psychologically possible to be irrational. In this case, however, the logical conflict underlying our irrationality does not hold between classically contradictory statements of the type '*p* and not-*p*.' Rather, it holds between the mutually exclusive epistemic states of 'being uncertain that *x*'—which Gordon claims is an essential component of 'fearing that *x*'—and 'knowing that *x*.' These states, however, are themselves mutually exclusive in two strong senses. Logically, 'being uncertain that *x*' *just means* (in part) 'not knowing that

x'. Psychologically, then, 'being uncertain that *x*' is by definition a mental state that immediately ceases to exist when we come to know that x. The outcome of Gordon's stricture **R** on 'fearing that' would thus seem to point to a stronger sense of impossibility: fearing that one will die is *psychologically* impossible. It is, simply, a mental state which humans cannot possess, analogous to impossible imaginings of square circles. But of course we certainly do seem to fear death even when such a fear is 'proven' to be impossible or irrational. This is just the partial point of Larkin's *Aubade*. Something must therefore give.

The most likely candidate is Gordon's claim that 'fearing that' *necessarily* involves uncertainty. Gordon derives this claim from nothing more than semantic intuition. That is, he simply takes the standard usage of 'S fears that *x*' to imply uncertainty. Despite the questionable methodology, though, Gordon does seem right here. As noted above, there is a definite contradictory ring to a subject's claim that they fear that their plane will crash even as they admit knowing that it has landed safely and rolled to a stop. It seems equally clear, however, that this contradictory ring is unique to the 'fear that' construction. Relying now on my own semantic intuitions, I would argue that there is no such contradictory ring to the claims 'I am afraid *of* dying' and 'I am certain I will die.' What then does Gordon make of 'of' constructions? If I cannot fear *that* I will die can I yet be afraid *of* dying?

Gordon's curative strategy here is to cast all gerundive nominalizations—i.e., 'of *x*ing' constructions—as transformations of embedded 'that' clauses. Thus according to Gordon the attribution 'She is afraid of slipping on the ice' is only a surface

transformation of the more explicit attribution 'She is afraid *that* she will slip on the ice.'[9] Gerundive nominalizations are thus to be construed as abbreviations of semantically equivalent 'that' clauses. Of course, some attributions contain nominal expressions that clearly are not abbreviations. 'Joe is afraid of the neighbour's dog,' as Gordon notes, is no abbreviation of any particular sentence. In these cases, though, Gordon argues that such constructions always *entail* some particular 'that' clause. Joe's fear *of* the neighbours dog, on Gordon's view, will always entail some sentence of the form 'Joe is afraid *that* the dog (will bite him; will harm him, etc.)' However the reduction is to be effected, then, Gordon is making an extremely strong point here. He sums up: "All fearing is "propositional,"...all fears are fears *that* something is (or: was, will be) the case."[10]

Returning to the fear of death, then, consider the following attributions of fear:

(3) S is afraid of dying.
(4) S is afraid of (his) death.

As (3) is a gerundive nominalization, recasting it in the form of an equivalent 'that' cause gives us

(3a) S is afraid that he will die.

As argued above, given (3a) and Gordon's conditions **R** for fear, S is seemingly cured of the fear of death. So long as S wishes not to die, and is certain that he must, he *cannot* fear death any more than he can picture a square circle.

---

[9] Ibid., p. 67.
[10] Ibid., p. 67.

Regarding (4), the nominal expression "his death" is not gerundive and so must imply some 'that' clause that specifies precisely what it is about his death that S fears. Obvious candidates include:

(4a) S is afraid that his death will be painful.
(4b) S is afraid that his death will be lonely.
(4c) S is afraid that the afterlife will be unpleasant.

The fear *of* one's own death is thus here instantly cured by an act of semantic transformation; it is not one's own death that is feared but rather some frightful state of affairs x attendant upon one's death. Of course, we still might fear that x so long as we wish x not to be the case and remain uncertain whether it will be so. Effecting a complete cure here involves some leg work. Two options present themselves. First, simply stop wishing x not to be the case. **R** fails to hold and the 'fear that x' is removed. If this option proves too difficult, however, a second remains. Upon discovering the value of x, simply ascertain for certain whether or not x will in fact occur. Again by **R** and *modus tollens*, certainty about x's obtaining renders impossible the 'fear that x.' The fear of one's own death, and whatever unpleasantness it might imply, is cured.

I doubt, however, that Gordon's cure will satisfy those among us for whom Larkin's *Aubade* rings true. Several reasons suggest themselves. First, it seems to me entirely possible that one might be afraid of death, yet be unable to say exactly what it is that they fear about it. Or more strongly, there might just not be anything in particular about death *that* they fear. In short, Gordon simply leaves unsupported his claim that every 'of' construction must imply some particular 'that' clause which describes more specifically exactly what is that is feared. The fear *of* death, therefore, could simply 'stand on its own,' unaffected by the logical problems that dissolve the fear *that* one will

die. Moreover, even if Gordon could prove his claim that 'of' clauses always imply some 'that' clause, the cure that this semantic transformation makes possible remains a hollow remedy. While I have argued—in the spirit of a *reductio ad absurdum*—that Gordon's stricture **R** on 'fearing that x,' coupled with the certainty that x, dissolves that fear, there lingers over the entire affair an air of sophistry. Supposing that **R** is a reasonable condition for the existence of fear, and that my argument is valid, then the fear of death seems to have been rendered as impossible as the imagining of a square circle. But clearly we do fear death, and even if Gordon is right and actually fearing death is an impossibility, then at the least the appearance of this fear must be explained. Of course, the reasonable assumption here, and my ultimate point in this discussion, is that the philosophical cure for the fear of death is merely an unwelcome artefact of Gordon's illicitly casting all emotion attributions as 'that' clauses. There is, however, a larger lesson to be learned.

(*Hyper-*) *Cognitive Theories of Emotion*

Gordon's approach to emotion would likely be classified by most as a "cognitive theory of emotion" because its central argument is that most emotions *necessarily* involve belief or knowledge. I would suggest, however, that Gordon's theory is also a prime example of what I will call 'hyper-cognitivism,' a philosophical stance that draws on the tradition of cognitivism initiated by Aristotle but distorts it in some systematic ways. In the following sections I will look at three of the most serious of these distortions: a methodological reliance on conceptual analysis to the exclusion of empirical research; a

tendency to argue that an emotion's cognitive element is what lends that emotion its particular identity; and a lack of recognition of the degree and significance of variation in the nature of emotion's cognitive elements. Before proceeding with this critique, however, I want to look briefly look at the notion of cognition and what qualifies a theory of emotion as cognitive.

Since Plato's recognition of the tripartite rational-emotional-appetitive soul, cognition has traditionally formed one third of the more modern triad of cognition, emotion, and motivation. Within this traditional view cognition is often ostensively defined by pointing to what Paul Griffiths calls the "traditional paradigms of "cognitive" processes."[11] Paradigms of cognition as *process* include, among others, perception, learning, memorization, and problem solving. Relatedly, these processes are typically seen as involving the manipulation of cognitive *states*; here *belief* is paradigmatic—at least within philosophical circles. Considered collectively, it is often further claimed that the paradigm processes definitive of cognition are *unified* in that they all centrally involve *information processing*, an identification I will discuss below. It is perhaps worth noting, though, that "cognition" is not a typical philosophical term. Philosophers have traditionally avoided speaking of "cognition" in general, retaining instead the classic vocabulary of belief, judgment, and perception.

As such, I'll frame my rough definition of what counts as a cognitive theory of emotion in similar terms. Loosely defined, a theory of emotion is cognitive if it claims that emotions are cognitively structured in that they are intimately 'tied' to

[11] Griffiths 1997, p. 25.

paradigmatically cognitive states like belief, judgment, or evaluation.[12] Here, the vagueness of 'tied' is intended to mark a significant variance in how different theories conceive of the relation between emotions and cognitions. Some, for example, claim the connection is *causal* in that certain cognitive states are necessary—or necessary and sufficient—conditions for the occurrence of emotion. Gordon's theory is an example here, given the belief condition he places on certain emotions. Appraisal theory, which I will discuss in a later chapter, is another example, though it conceives of the cognitive components of emotion somewhat differently than Gordon. Other theories claim the relationship is one of identity: emotions *just are* a special class of cognitive states. Robert Solomon's theory, in which he equates emotions with evaluative judgments, is a classic example of this type.[13] Aristotle's theory of emotion might also count here, though he identifies emotions with a complex of elements, only one of which is cognitive.

What makes such theories 'cognitive,' despite their differences, can be sharpened by noting the few theories that are most commonly described as non-cognitive. Certainly any theory that held Dylar out to be a real possibility would count as non-cognitive since it would allow that an emotion could be wholly dissolved without *any* change in a

---

[12] Cf. William Lyons' definition: "In general a cognitivist theory of emotion is one that makes some aspect of thought, usually a belief, central to the concept of emotion and, at least in some cognitive theories, essential to distinguishing different emotions from one another" (Lyons 1980, p. 33). My own definition is modelled on Lyons' but I have tried to avoid the overly inclusive modifier "some aspect of thought" which renders almost every theory of emotion a cognitive one.

[13] "...my embarrassment *is* my judgement to the effect that I am in an exceedingly awkward situation. My shame *is* my judgement to the effect that I am responsible for an untoward situation or incident. My sadness, my sorrow, and my grief *are* judgements of various severity to the effect that I have suffered a loss. An emotion is an evaluative (or a "normative") judgement, a judgement about my situation and about myself and/or about all other people. Needless to say, this is not the usual portrait of the emotions. The emotions are usually thought to be *consequent* to judgements, perhaps a slightly delayed reaction to their import, but not the judgements themselves" (Solomon 1993, p. 126).

subject's cognitive economy. It thus follows that the obviated emotion must not have depended upon any form of cognition for its existence.

A somewhat less fabulous example is William James' equation of emotion with the consciousness of particular physiological changes. James noted that the commonsense view of emotion was that the perception of some object or fact caused an emotion and that this emotion then gave rise to some physical expression: we see a bear, grow frightened and run. James reversed this formula: "My theory, on the contrary, is that *the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur IS the emotion.*"[14] James' theory is thus best described, following de Sousa, as *physiological*, since it is the occurrence of bodily changes that are necessary for the occurrence of an emotion: "Without the bodily states following on the perception, the latter would be purely cognitive in form, pale colorless, destitute of emotional warmth."[15] Of course, perception of the exciting fact, which for James is an instance of cognition, is also necessary for the occurrence of emotion. James thus recognizes a role for cognition in emotion, though a rather mundane one; it is the bodily change alone that lend particular emotions their identity and character.

A slightly different form of non-cognitive theory can be reconstructed from Hume's work on the passions. While even a cursory examination of Hume's work shows that he recognizes a central role for belief and other 'cognitive factors' in emotion, some commentators—such as Anthony Kenny—have fixed on the following quote to show that Hume held a somewhat bizarre non-cognitive theory of emotion:

---

[14] James 1893, p. 449.
[15] Ibid., p. 450.

> A passion is an original existence, or, of you will, modification of existence, and contains not any representative quality, which renders it a copy of any other existence or modification. When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high.[16]

On the view encapsulated here, emotions are a curious form of *sui generis* mental states that are distinguished, in part, by their being *non-representational*. Cognitive states, conversely, are traditionally thought to be representational states *par excellence*; beliefs, judgments, memories and such are always *about* some state of affairs. They *refer* to those states of affairs—in a way we have yet to understand. Humean emotions, however, apparently do not refer and hence are not cognitive states. Indeed, on at least one view of what marks off the realm of the mental, Humean emotions would not even count as mental states since on this view "mental states *are* representational states."[17]

Lack of reference or representational capacity can also mark out a theory as non-cognitive in a slightly less suspect way. Consider, for example, Georges Rey's view of emotion—it is too undeveloped to be called a theory—that is in one sense classically cognitive. He offers as a "tentative hypothesis" the following characterization of emotions. Typically , they are:

> complex states involving nomological interactions between cognitions, qualitative states, and physiological states....For example, a specific cognition, or constellation of cognitions, might be linked nomologically to specific qualitative and physiological states, and so forth; a given emotion might be regarded as some commonly occurring segment of just such a sequence.[18]

---

[16] *A Treatise Of Human Nature*, II.I.vi. See also III.I.i. For an account of Hume as cognitive theorist, see Davidson 1980, pp. 277-290.
[17] Sterelny 1990, p. 19.
[18] Rey 1980, p. 188.

So expressed Rey's view is strongly cognitive in that it sees particular cognitions playing a central causal role in the production of emotion. It is a cognitive view even while it allows that emotions also involve important non-cognitive factors—there are few purely cognitive theories in the sense of claiming that all emotions are wholly composed of cognitive factors.[19] Rey emphasizes, for example, the significant role that the hormonal system plays in the production and maintenance of emotions. Importantly, though, the fact that emotions involve these different elements leads Rey to deny that emotions should be conceived of as simple *states*. He argues instead that they are best understood as *processes*, i.e., temporally ordered complexes of disparate elements. Depression about the collapse of my career, for example, is identified as "the sequence beginning with the belief that one's career has indeed collapsed, the quite strong preference that it hadn't, a consequent depletion of norepinephrine, the effects of that depletion upon the nervous system, consequent further changes in cognition [and so forth]."[20]

What makes Rey's view interesting here is the conclusion that he draws from the fact that emotion involves an interaction of cognitive and non-cognitive systems. He notes: "We so far have no reason whatever to believe that the actions and the interactions of the hormonal with the cognitive system involve any rational relations among representations at all. The relations that do obtain appear to be *merely causal* ones."[21] Here, Rey is implicitly relying on a common conception of cognition as information processing, one expressed by Kim Sterelny: "...cognition consists of operations on states of our central nervous system, states which have meaning or

---

[19] Solomon's theory is arguably a pure cognitive theory.
[20] Rey 1980, p. 188.

content."[22] States of the hormonal system, however, presumably lack meaning or content; they are non-representational and as such can't enter into rational transactions (but see below). Considered as *entire processes*, then, it follows that on Rey's view emotions are non-cognitive in the sense that even though they *involve* cognitive states, the entire process that we identify as the emotion is non-cognitive in that it does not *proceed* through the rational transformation of representational states.

Of course, Rey's view relies on a sharp distinction between cognitive and hormonal systems, one which may in reality be unsustainable. Ronald de Sousa, for example, notes that cognition is susceptible to various hormonally induced effects.[23] Moreover, some neurotransmitters used in the nervous system are similar in chemical structure to hormones, and both neurotransmitters and hormones work in the same way by spreading diffusely rather than by traveling along preset channels. These considerations lead de Sousa to draw a looser distinction between slow (chemical-hormonal) and fast (nervous system) information processing systems in the body.[24] Rey, however, also recognizes hormonal effects on cognition. His account of depression, for example, recognizes that a depletion of norepinephrine will affect the nervous system and subsequently cause further changes in cognition. Even while recognizing these effects though, Rey retains a sharp hormonal-cognitive distinction, presumably on the basis that the effects are, in his terms, "strictly causal," since hormonal states are non-representational and so can't interact rationally with the representational states of the

---

[21] Ibid., p. 191.
[22] Sterelny 1990, p. 34.
[23] De Sousa 1990, p.69.
[24] Ibid.

nervous system . Here, however, there is more fuzziness, since de Sousa claims that the chemical-hormonal system *is* an information bearing system, albeit one that covaries unreliably with the environment.

A similar fuzziness about the nature of cognition lies at the heart of an important debate in the early 1980's between Robert Zajonc and Richard Lazarus, a clash that I will discuss in more detail in the next chapter.[25] The debate ostensibly revolved around whether or not cognition was a necessary precondition for basic emotional responses like preference formation. Zajonc cited a number of empirical studies that suggested that such responses proceeded independently of cognition. In response, Lazarus accepted what he saw as Zajonc's general empirical point, allowing that "we do not have to have complete information to react emotionally."[26] In fact, Lazarus had proved this himself in a controversial early experiment. He had shown that by pairing neutral stimuli—nonsense syllables—with an electric shock, subjects later responded to those stimuli with a galvanic skin response even when the stimuli were re-presented under degraded viewing conditions that precluded conscious perception and recognition. However, while agreeing with the general empirical claim, Lazarus denied Zajonc's conclusion, insisting instead that emotion always involved cognition *properly understood*. Zajonc merely had too narrow a conception of cognition. Specifically, Lazarus argued that Zajonc seemed to mistakenly claim cognition must always be deliberate, rational, and conscious. Once the definition of cognition was properly expanded to include non-deliberate, unconscious states and processes, Zajonc's argument, while factually accurate, obviously drew a

[25] The classic first papers are Zajonc 1980 and Lazarus 1982.
[26] Lazarus 1982, p. 1021.

mistaken conclusion. Zajonc, however, actually had a rather inclusive definition of cognition. He agreed with Lazarus that cognition "need not be deliberate, rational, or conscious." Cognition, for Zajonc, need only involve "some form of transformation of a present or past sensory input...according to a more or less fixed code."[27] It should be clear why it seemed to many that the debate was largely a terminological scuffle.

I won't try here to sort out the issues this debate raises. Rather, I want to close this section now by drawing some lessons from this brief discussion of cognition and cognitive theories of emotion, the first of which should be obvious: cognition is a poorly defined concept. The tendency to define cognition through ostension does little to sharpen the concept because the processes and states pointed to as exemplars are themselves ill-defined and not well understood. Similarly, the current tendency to equate cognition with information processing, while somewhat more substantive as a definition, still glosses over important differences. For example, the psychologist Carrol Izard usefully notes that DNA molecules process information; we would not want to view this, however, as an instance of cognition.[28] Having said this, however, the debate between Lazarus and Zajonc and the minor clash between Rey and de Sousa amply illustrate the second lesson to be drawn here: it is less important to come to a precise definition of cognition than it is to get straight on the details of the processes involved in emotion.[29] It

---

[27] Zajonc 1984, p. 261.

[28] Izard 1993, p. 70. Izard proposes that information processing be viewed as a continuum along which we recognise four basic types: cellular, organismic, biopsychological, and cognitive.

[29] In fact, it is entirely possible that in so doing the concept of cognition might turn out to be essentially useless. My reasoning here follows the similar argument made by Griffiths about the concept of emotion (1997, p. 14). The diverse processes typically collected under the category "cognition"— perception, learning, problem solving, etc.,—are presumed to be alike in important ways. Cognition, like emotion, is "meant to be a kind of psychological process that underlies a certain range of human behaviours" (ibid).

is uncontroversial now that emotions are subserved by a range of processes, some clearly cognitive, on any definition of the term, some debatably cognitive, and some clearly non-cognitive. Any complete understanding of emotion must therefore take this complexity into account. In particular, theories of emotion that focus on its cognitive aspects must ultimately connect those with emotion's non-cognitive elements. This might seem obvious, but as the following section shows, it is a lesson lost on theorists tending towards hyper-cognitivism.

Conceptual Analysis

Hyper-cognitivist theories are first distinguished by their near complete methodological dependence upon conceptual analysis and related lack of concern for empirical research. Conceptual analysis is in part a process defined by its objects of study. In contrast to the empirical sciences, which take as their focus the real physical particulars of the natural world, conceptual analysis instead studies the language and concepts we use to talk about the natural world. In the realm of philosophical psychology these are typically the everyday first person avowals and third person attributions of mental states that constitute the bulk of our common psychological discourse: 'I am afraid of dogs,' 'He believes he was cheated,' and so on. Beyond this initial defining aspect, however, conceptual analysis divides further two distinct forms, one having as its goal the definition of the central terms in its realm of study, the other having as its goal

---

As we come to understand these paradigmatically cognitive processes in detail, however, we might fail to find enough important common elements to justify grouping those processes under a single heading.

the explication of the implicit theory that underwrites common understanding of that realm.

In its former aspect conceptual analysis is now commonly seen as a process best avoided, especially when the definitions it seeks to provide are couched in terms of necessary and sufficient conditions.[30] In the light of seemingly endless failures the task of providing necessary and sufficient conditions for virtually anything seems a fool's game. We can call this form 'reductive' conceptual analysis as its central purpose is to provide definitions in which the terms of the *definiens* are more basic or perspicuous than the *definiendum*. In its second form, however, conceptual analysis is not essentially reductive. Here conceptual analysis aims at the explication of the implicit theory that governs our application of conceptual terms within a given realm. Conceptual analysis within philosophical psychology thus takes the form of an explication of the folk psychological theory governing our use of belief, desire, and emotion terms. 'Definitions' provided in this tradition are thus best understood as being *hypotheses about the actual nature of the definiendum*. Gordon, for example, reads much of the history of philosophy of emotion as being conceptual analysis of this second form:

> Remarkably many of the major classical philosophers took it as a major challenge to their analytical skills to attempt definitions of the various emotions...Aristotle in the *Rhetoric*, Descartes in *The Passions of the Soul*, Hobbes in the *Leviathan*, Spinoza in his *Ethics*, and Hume in *A Treatise on Human Nature*. What they were doing in their defining, I suggest, was to make explicit the elaborate commonsense theory [of emotions].[31]

---

[30] Stich 1983, p. 77.

[31] Gordon 1987, pp. 9-10. Gordon similarly reads Grice's theory of speaker meaning as being "a hypothesis that certain *inferential processes* go in the planning of speech behaviour" (ibid., p. 11). Gordon locates himself squarely within this long tradition. What makes him unique, however, is that he builds his theory of emotion upon an analysis of the semantics of folk-psychological attributions of emotion that take the narrow form 'S emotes *that*.' As I have argued above, this narrowness sometimes leads to unacceptable conclusions. I say 'sometimes' here only because the particular problem I identified in

These definitions, in Gordon's words, are thus intended "to tell us something about ourselves."[32] We should understand, for example, Aristotle's definition of anger as a hypothesis that an angry subject *is actually possessed of* 1) a *desire* for revenge, accompanied by 2) a painful *hedonic element*, where these two aspects are *caused* by the *judgement* that one has been slighted.[33]

Such production of hypotheses through the analysis of folk theory is in itself a benign process. There is nothing inherently wrong with simply creating hypotheses. Conceptual analysis of this second form, however, as put to use by recent philosophers of emotion, has earned a bad name by marrying itself to an outmoded semantics that takes the reference of a term to be completely fixed by the rules governing the application of that term in common discourse.

This at least is the view of Paul Griffiths. In a lengthy critique of the use of conceptual analysis in the philosophical study of emotions Griffiths argues that most modern philosophers of emotion have ignored the turn from descriptive to causal theories of meaning:

> Propositional attitude theorists think conceptual analysis is the only tool they need to investigate emotions because they accept, explicitly or implicitly, a Wittgensteinian distinction between the "criteria" which logically define a mental state and the inessential "symptoms" that can be studied empirically. Mental states are *defined* by the rules which ordinary speakers use when applying mental state terms.[34]

Gordon's theory could likely be avoided by a similar theory which considered a wider class of third-person attributions and first-person avowals. Indeed, as we have seen, most of the philosophical theories Gordon mentions above do precisely this. Neither Descartes nor Aristotle restrict themselves to particular grammatical forms of attribution. Moreover, most current psychological semantically oriented theories are similarly unrestrictive in their range of analysis.

[32] Gordon 1987, p. 14.

[33] *Rhetoric*, II.ii.i.

[34] Griffiths, 1997, p.23.

In so far as conceptual analysis amounts to the explication of folk theories, once we reject the claim that a concept's meaning is exhausted by the current linguistic rules governing its application, we see that conceptual analysis is radically limited. *All* that it can do is make explicit current beliefs about a concept. It cannot tell us whether those beliefs are true. Aristotle's definition of anger, for example, might be an adequate analysis of what people generally believe about anger, but there is no particular reason why we should accept it as a truth about the psychological processes and mechanisms that actually produce anger.

While I agree with Griffiths' general criticism, I would suggest that conceptual analysis still has a place in the philosophy of emotion. Pursued in its more benign form as the explication of implicit theory, conceptual analysis has value as a ground for the generation of *testable* hypotheses. So long as these hypotheses are subsequently put to the empirical test there seems no good reason for a wholesale abandonment of conceptual analysis. Stephen Stich, for example, suggests that conceptual analysis should align itself with the computational study of cognition as a means of providing general frameworks used to guide and constrain the construction and implementation of computational models: "...philosophical [analysis] can be viewed as giving a rather coarse-grained discursive characterisation...for the more detailed program that the cognitive simulator is trying to write."[35] This has actually already occurred within the field of emotion studies. There are a number of computational models of emotional processes that are founded upon conceptual models of emotion developed through an analysis of common emotional

[35] Stich 1983, p. 77.

discourse. The cognitive theories being simulated in these programs, however, have been drawn almost exclusively from work in experimental psychology.[36]

Cognition and the Individuation of Emotions

In addition to its dependence upon conceptual analysis, hyper-cognitivism is distinguished by a second, closely related feature: the claim that the cognitive factors operative in the production of an emotion are the central determinant of the identity of that emotion. I have already discussed two of the earliest and most influential forms of this claim. As previously noted, both Aristotle and Descartes identified a discrete set of *dimensions of evaluation* along which subjects judged a situation or stimulus in the course of experiencing an emotion. Recall, for example, Aristotle's account of anger's defining cognition: the judgement *type* 'I have been slighted.' This 'judgement' was, for Aristotle, only a formal placeholder, i.e., an evaluative category under which fall the specific judgement tokens that cause particular, concrete instances of anger. More than simply causing anger, however, Aristotle also held that formal judgements served to differentiate emotions: anger is *by definition* a state caused by judgements of the type 'I have been slighted.'

This notion of a defining, formal cognitive structure that both elicits and differentiates our specific emotions has been carried through into modern philosophical treatments of emotion and is most strongly expressed in the claim that specific emotion types—anger, fear, envy, etc.,—are defined or individuated by a unique *formal object.*

---

[36] For comprehensive surveys of computational models of emotion see Picard 1997 and Pfeifer 1988.

So expressed, this claim dates to Anthony Kenny's influential *Action, Emotion and Will*. Kenny defines formal objects as follows: "The formal object of ϕing is the object under that description which must apply to it if it is to be possible to ϕ it. If only what is P can be ϕd, then "thing which is P" gives the formal object of ϕing."[37] For Kenny the import of this notion is that the assignment of a formal object to some action constrains and defines the nature of that action. For example, "one's own spouse" is the formal object of the act of divorce. If that description does not apply to the direct object of the verb describing the action then the action cannot be performed: I can only divorce someone who is my spouse.[38] In the case of verbs denoting psychological acts, however, the particular defining description need not apply in actuality to the object, but rather must only be believed to apply: only things which are actually wet can be dried, "but something which is merely believed to be an insult may provoke anger."[39] The essential defining feature of an emotion is thus not any particular property the object of that emotion has, but is rather the content of that description which the subject judges to apply to that object.

Of particular importance here is Kenny's view that such judgements and the emotion types they define are related as a matter of logic: "each of the emotions is appropriate—logically, and not just morally appropriate—only to certain restricted [formal] objects."[40] Precisely how Kenny understands the nature of this relationship, or

---

[37] Kenny 1963, p. 189.
[38] Recall that Aristotle also made this point.
[39] Kenny 1963, p. 194.
[40] Ibid., p. 192. Kenny often slips uncritically between speaking of (1) a logical relationship between emotions and their objects – understood as concrete particulars which are the target or focus of the emotion, and (2) a logical relationship between emotions and their formal objects. It is clear, though, that

how it originates, is not entirely clear. He never speaks explicitly to either issue. From his general account, however, we may reconstruct his most likely answer to the question of precisely *what* the logical relationship between emotion types and formal objects amounts to.

Kenny seems to base his claim for this relationship largely upon the principle that our understanding of particular emotions is immune to revision stemming from new-found empirical knowledge.[41] To see Kenny's point here, consider his example of a man who claims to fear winning the lottery: "If we can elicit from him only descriptions of the good aspects of the situation, then we cannot understand why he reports his emotion as fear and not as hope."[42] Faced with such a subject we must continue to look for extenuating circumstances that might explain his fear: perhaps he worries he will become the target of thieves. Failing in this regard we must simply assume the subject doesn't understand the meaning of "fear." The point here is that regardless of which option we choose there is no further empirically discoverable fact about the person that would convince us to accept that he actually fears something he judges to be wholly good. Thus the emotion type 'fear' might be claimed to be logically connected to its defining formal object in the sense that no empirically discoverable fact would lead us to accept some state S as an instance of 'fearing x' unless S was founded on a subject's assenting to some appropriately negative description of x.

---

he views the logical relationship to hold between the latter pair since virtually any particular object can be the focus of any emotion so long as the right sort of belief is held about it.

[41] For Kenny's explicit claim that empirical psychology can tell us nothing *essential* about the nature of our emotions see p. 51.

[42] Ibid., p. 180.

Of course, this connection between logical necessity and immunity to revision is now widely rejected following Quine's attack on the analytic/synthetic distinction.[43] Subsequent philosophers in the hypercognitivist tradition, however, have still followed Kenny in his claim that emotions are logically connected to their eliciting cognitions, though for slightly different reasons. Later claims for a logically individuating connection between emotion and cognition have tended instead to rest upon observations of the following sort. It appears inconsistent for a subject to claim they are angry that their car was stolen yet deny that they believe their car was stolen. This apparent inconsistency in turn seems to lead directly to the conclusion that the 'emotion that *p*' and the 'belief that *p*' are logically related, where this logical relation consists in the fact that the denial of this relation is inconsistent.[44] It seems, then, that 'emoting that *p*' requires the belief that *p*.[45] This is the sort of reasoning, for example, that Gordon implicitly relies upon to justify his belief condition BC on factive emotions.[46]

However compelling one finds these arguments their relevance to the claim that emotions are individuated by formal objects trades on a fundamental ambiguity. It is one thing to note that emoting that *p* in some sense necessarily requires believing that *p*. The required belief that *p*, however, is not a formal object. As Kenny and others have intended the notion a formal object is a 'category' or 'type' description that the subject judges, via some *tokened* mental state, to apply to the object of the emotion, and it is the identity or content of *this* category that determines the identity of the emotion elicited by

---

[43] Quine 1951.

[44] In much the same way, the apparent inconsistency of denying the truth of statements like "All bachelors are unmarried males" was a central intuition underlying the claim that such statements were analytic (Churchland 1979, p. 46).

[45] See e.g., Davidson 1976, p. 289.

the tokened mental state. Straightforwardly factual beliefs of the sort "my car was stolen" cannot identify or individuate an emotion since a given belief of this type can support any number of emotions. Believing my car stolen I *might* become angry, but I might equally well remain indifferent, or I might even be happy that it was stolen since I can collect some much needed insurance. Thus for Kenny the important *defining* logical relation is not between my anger that my car was stolen and my belief that it was stolen, but between my anger and—following Aristotle—the judgement that the theft of my car is an instance of being slighted.

Once the terms of this relation are more clearly defined the logical necessity that seems to unproblematically hold between being angry that p and believing that p becomes largely irrelevant. Consider:

(1) I am angry that my car was stolen
(2) I don't believe that my car was stolen.

(3) I am angry that my car was stolen.
(4) I don't believe that having my car stolen was an instance of being slighted.

Unlike the first set of sentences, which seem *prima facie* inconsistent, the second set do not. We can easily imagine someone becoming angry that their car was stolen while denying they see it as a slight. Perhaps they see it *only* as an inconvenience. Given this current understanding of what the logical connection might amount to, it does not then clearly follow that a subject's anger that their car was stolen is logically connected to the judgement that having one's car stolen is an instance of slight. Of course, a proponent of formal objects might argue that all that is required to establish the required logical connection is to come up with a more adequate characterisation of anger's defining

---

[46] See e.g., Gordon 1987, p. 36.

formal object. The reason we do not find (3) and (4) inconsistent might just be because the description contained in (4)—the description of car theft as "an instance of slight"—doesn't really capture the 'identity' or 'content' of anger's defining formal object. There might be, then, a more adequate formulation of this description that when substituted into (4) would render it clearly inconsistent with (3) and thereby establish the desired logical relationship between that description and the emotion type 'anger.' This response, however, raises a major difficulty facing the notion of formal objects: the problem of convincingly formulating the actual content of the specific formal objects that supposedly individuate our various emotion types.

The difficulties of this project are especially apparent in those later philosophers who, while adopting Kenny's general theoretical structure, have turned their attention away from the apparently logical nature of the emotion-cognition relationship and focused instead on understanding the wider implications of this relationship. Following Kenny, two major modern works on emotion, William Lyons' *Emotion* and Ronald de Sousa's *The Rationality of Emotion*, make similar but more extensive use of the notion of a formal object. Their respective definitions run as follows:

> (Lyons) The formal object of an emotion seems to be the *evaluative* category under which the appraisal or evaluation of a particular object...falls on a particular occasion. Indeed the fact that the particular evaluation or appraisal falls under the general evaluative category associated with an emotion as part of its definition or concept is our ultimate licence for saying that this emotional state is of such and such an emotion.[47]

> (de Sousa) For each emotion, there is a second-order property that must be implicitly ascribed to the [emotion's object] if the emotion is to be intelligible. This essential element in the structure of each emotion is its *formal object*....The

[47] Lyons 1980, p. 100.

> specific formal object associated with a given emotion is essential to the definition of that particular emotion.[48]

Both definitions express the same essential idea as Kenny's: an abstract evaluative category *C* defines a particular concrete emotion *e* as being of type *E* in so far as the actual cognition *c* which caused *e* falls under the category *C*. So, again, if we accept Aristotle's analysis, what makes my current emotional state an instance of anger is that it arose from my judgement that your behaviour toward me just now was contemptful, i.e., I ascribed to your behaviour the second-order property of being contemptful. And, given that contempt is one of the "three species of slight," my particular tokened judgement in this case falls under anger's defining evaluative category of judging that I have been slighted. Lyons offers a similar account of the differentiation of fear, love, and grief:

> An emotional state is labelled as 'fear' rather than 'love' or 'grief' because the feelings, physiological changes and desires—and the ensuing behaviour if any—which form the state are believed to be the result of an evaluation that something is dangerous rather than that it is appealing...which would be the evaluation typical of love, or that it is a grave loss or misfortune, which would be the evaluation typical of grief.[49]

Even abandoning this concern with showing the connection to be a logical one, however, the notion of an individuating formal object still faces a number of problems.

Lyons himself considers the first objection: "It might be objected...that 'the dangerous' does not really capture the formal object of fear because one can evaluate something as dangerous and be, say, excited rather than afraid."[50] The point here is that, considered as a general evaluative category, 'dangerous' is too vague to differentiate

---

[48] de Sousa 1990, p. 122. The notion of a defining formal object is especially important to de Sousa's project as it provides the possibility of establishing a criterion of 'correctness' for emotions. Very roughly, an emotion is *appropriate* if the object of that emotion actually possesses the second-order property implicitly ascribed to it (ibid).

[49] Lyons 1980, p. 100.

between fear and excitement. Similarly, evaluations of *x* as 'appealing' wouldn't distinguish 'loving *x*' from merely 'liking *x*', nor would evaluations of *x* as being a misfortune differentiate being 'sad that *x*' from 'grieving that *x*.' In each case it seems that the evaluative category needs 'sharpening' in order to serve as a differentia for its associated emotion. Lyons notes:

> Fully to separate the evaluative aspects of fear and excitement one would have to go into more detail. At the least one would have to spell out the evaluation of the object of fear as not merely 'dangerous' but 'disagreeably so' as well, for the person who is excited, even if by danger, cannot claim to find the danger disagreeable.[51]

It might seem then that all that is needed is further analysis to produce an evaluative category sufficiently precise to distinguish 'like' from 'love' and 'fear' from 'excitement.' Moreover, it might be objected that fear and excitement, like love and like, actually are in a broad sense the same emotions, perhaps differing only in degree. As such their sharing the same very general formal object is unproblematic. To answer these counter-objections it will be helpful to consider a case in which Lyons actually provides a detailed analysis of the defining evaluations of distinctly different emotions.

Lyons notes that it seems reasonable to suppose that the evaluation 'Ashkenazy is a fine pianist' could be central to the opposite emotions of envy and admiration. At a very general level both emotions seem to necessarily involve a positive evaluation of the pianist Ashkenazy. As such, this seems to speak against the claim that a simple positive evaluation defines either emotion. What is needed, Lyons argues, is a fuller analysis of the different emotions' defining cognitions.

---

[50] Ibid., p. 101.
[51] Ibid.

Lyons' subsequent analysis identifies four distinct elements within envy's definitive evaluation:

1. The evaluation must note a gap between the subject experiencing the emotion and the object of that emotion; in regard to his talent alone, I will not envy a pianist I believe to be my equal.
2. The subject must further see himself as falling on the inferior side of the gap; in respect to skill alone, I will not envy a pianist I judge to be inferior to me.
3. The object of the evaluation must be something in which the subject has an interest in; I will not envy another pianist's skill if playing skilfully does not matter to me.

These further evaluations go some way toward separating envy from admiration since evaluations of this sort will in most cases be sufficient to trigger envy while discouraging admiration. Lyons notes, however, that it is not inconceivable to suppose someone might be a pianist, thus satisfying the third requirement, while recognising Ashkenazy's greater skill, thus satisfying condition 1 and 2, while still admiring Ashkenazy. A fourth requirement is thus need that will definitively separate problematic cases into instances of envy or admiration:

4. I must evaluate my being inferior as "being displeasing or not to [my] liking; "I could not say that I admired Ashkenazy's skill and that part of this admiration involved being displeased by the realisation that my playing of the piano was

> vastly inferior to his....[b]eing displeased is part of the concept of envy but not part of the concept of admiration."[52]

Lyons argues that this further analysis provides a more comprehensive picture of envy's defining evaluation, one that adequately describes how envy differs from admiration. While both emotions rest in part upon a positive evaluation of some aspect of the emotion's object this simple evaluation does not alone suffice to define either emotion. It is only when this evaluation occurs in concert with the above four conditions that envy is produced.

It should be obvious, however, that there is a significant problem in Lyons' account. Simply put, Lyon's third and fourth conditions are highly problematic. 'Being interested in' and 'being displeased by' seem themselves to be something like emotions! At the very least, neither condition is an unproblematically, purely cognitive evaluation. Lyons thus owes some explanation of why we should consider these conditions to be on a par with factual judgements of the sort 'he is a better pianist than I.' Short of this, Lyons is simply illicitly appealing to non-cognitive, quasi-emotional elements in his attempt to give an account of the purely *cognitive* aspect of emotion that he holds to be definitive of emotion types.

This particular move, in fact, occurs frequently throughout the cognitive approach to emotion. Consider, for example, Robert Solomon's reply to what might be called the problem of unemotional evaluation. Any theory that ties emotions to particular evaluations or judgements is faced with the fact that it is virtually always possible for a person to make the judgements that some theory has tied to a particular emotion, yet not

---

[52] Ibid., p. 83.

experience that or any other emotion. I know, for example, that driving on the highway is dangerous, but I am not frightened by the prospect. Solomon notes:

> One might make a judgement—or even much of a set of judgements—in an impersonal and uninvolved way, without caring one way or the other. But an emotional (set of) judgement(s) is necessarily *personal* and *involved*. Compare "What he said to me was offensive" (but I don't care what he thinks) and "He offended me!" Only the latter is constitutive of anger."[53]

The point here is that what counts in the formation and identity of anger is not just the content of relative judgements but also depends upon how those judgements are 'held.' Like Lyons, Solomon here appeals to an unanalysed, non-cognitive element—call it 'personal involvement'—in the course of developing a supposedly 'pure' cognitive theory of emotion, i.e., one that rather straightforwardly identifies emotions with judgements. Similarly, recall that in his explication of the makeup of particular emotions Gordon adds 'wishing,' in various forms, to the cognitive elements belief and knowledge: "When *S* is angry about the fact that *p*, *S* believes that *p* and wishes it not to be the case that *p*."[54] A similar move is made by de Sousa in his discussion of the necessary role of motivation in his fuller account of the relation between emotions and their objects.[55]

---

[53] Solomon 1980, p. 276.

[54] Gordon 1987, p. 68.

[55] To fully specify the identity of a particular emotion, de Sousa argues that in addition to knowing the relevant formal object we must also know its target—"that real object, if any, *at* which the emotion is directed"—and its 'motivating aspect' (de Sousa 1990, p. 123). The motivating aspect of an emotion is the particular property of the emotion's target that causes the emotion and is rationally related to the emotion in that it constitutes an intelligible rationalisation for the emotion. Wendy, for example, despises Bernie. She thinks it is because of his poor taste, but in reality it is because he is Jewish. Here, Bernie is the target of Wendy's contempt. The motivating aspect here, however, is Bernie's ethnicity, as it is his being Jewish that caused Wendy to despise him, and it his being Jewish that, when coupled with Wendy's anti-Semitism, makes intelligible her contempt. De Sousa here leaves unemphasized the already emotional aspect of motivation. Motivation might belong, as de Sousa notes, to the sphere of rational discourse in so far as it rationalises and makes intelligible our particular emotions. Motivation, however, is not a purely cognitive or rational concept. An appeal to motivation, like Lyons' appeal to interest, and Solomon's appeal to personal, involved judgement, is an appeal to something already like an emotional state.

Of course, there is nothing inherently wrong with modifying a cognitive theory by including appeals to factors that are themselves not purely cognitive like interest, personal involvement, motivation, wishing, and desire. The prevalence of this form of modification is perhaps just evidence that any attempt to account for the individuation of emotion in purely cognitive terms is untenable. In practice, though, the modified theories noted above face some difficult general problems.

First, and most significantly, not all emotions need involve the various additional elements the above theorists claim. Consider, for example, Gordon's analysis of sadness. Gordon—tellingly, I believe—nowhere explicitly discusses sadness and its cognates, but we may easily enough construct an analysis in line with Gordon's general account.[56] Sadness is for Gordon a factive emotion since it first requires that the subject be certain about the truth of the belief that grounds their sadness; it would make no sense, presumably, for me to be sad that my grandfather has died while being uncertain that he has died. As a second general condition, however, I must also wish that it not be the case that he has died. I would argue, however, that this is simply mistaken. Even as I feel sorrow over his passing I could, without pain of contradiction, have wished for his death as a merciful end to his painful suffering. More generally, it simply seems unjustified to claim that *all* emotions, or all instances of a particular emotion type, must involve some form of wishing. What, for example, am I wishing for when I am disgusted by a friend's raging promiscuity? Gordon claims that 'disgust that *p*' requires the wish that not-*p*, but can I not be disgusted at my friend's behaviour while not caring one way or the other if it

[56] I base my analysis here on Gordon 1987, p. 32.

continues—"I don't care what you do, just don't tell me about it anymore!"?[57] In addition, as Paul Griffiths points out, positive emotions resulting from satisfied desires pose a similar problem.[58] I can be happy that I won the lottery even after my success has led me to stop wishing for a win. Or I might not have known I was entered in the lottery and thus never had the occasion to wish for a win

Similarly, it is not clear, as Solomon claims, that all judgements involved in emotion are always held in a 'personal' or 'involved' way, whatever such terms might mean. If, following de Sousa, we understand Solomon as meaning that emotional judgements are those held "with particular intensity," then our milder emotions prove difficult to explain.[59] Some emotions, moreover, are distinguished by their lack of intensity, indeed, by their lack of 'emotionality.' Cool anger and quiet sadness are examples here. It is, however, fruitless to pursue this particular criticism too far because Solomon never spells out in detail exactly what 'personal involvement' amounts to.

This last point, however, immediately raises the second main difficulty with modified cognitive theories of the sort outlined above. In general, the additional non-cognitive elements posited in these theories are often left unanalysed and unexplained. This is a particularly glaring omission because these elements are intended to do the essential work of transforming non-emotional beliefs, judgements, evaluations and such, *into* emotions. And this in addition to helping individuate emotions from one another. What is supposed to make the difference between merely knowing that you betrayed my

---

[57] Here I do wish something, that I not be told, but the object of my wish is clearly distinct from the object of my disgust.
[58] Griffiths 1997, p. 31.
[59] De Sousa 1987, p. 41

confidence and being angry that you did so? *Wishing* that you had not. Yet, like Solomon's silence on personal involvement, Gordon says little about what wishing amounts to.

Of course, other analyses of emotion that follow the general strategy of 'sharpening' definitive formal objects by analysing them into sets of judgements might somehow avoid the problem of illicitly appealing to unanalysed quasi-emotional elements, even though it is difficult to see how they might then answer the problem of unemotional evaluations. I would argue, however, that even if this were somehow accomplished the general approach embodied in the theory of formal objects faces another significant, purely formal difficulty. In short, it seems impossible to achieve a balance between (1) characterising an emotion's defining formal object *precisely* enough so that it will suitably differentiate between related but different emotions, and (2) characterising the formal object *generally* enough that it does not *exclude* an emotion from its obviously correct category. In the first instance, as an evaluative category or formal object becomes more precise it also becomes progressively more exclusive. Consider the following example. S claims that he is afraid that his girlfriend will not accept his proposal. According to both Lyons and de Sousa, if S is in a state of fear than that state must have been caused by a concrete cognition—an appraisal, judgement, belief, etc.,—where that cognition is of the *type* 'is dangerous,' or to use Lyon's more precise category, 'is disagreeably dangerous.'[60] When we ask S, however, why he is afraid, he answers only that he thinks his beloved's failure to accept 'would be a disaster.'

[60] Or to use de Sousa's terminology, S must have implicitly ascribed the property 'dangerous' to the failed proposal.

Pressed further he says it would be humiliating and ruin his carefully laid plans for starting a family. If we accept S's testimony at face value we are now presented with three articulated, concrete conscious judgements that seem to have caused his current state of fear. The important question now is whether any of these judgements are of the type 'is dangerous' If they are not, according to Lyons, then S cannot be in a state of fear. I would argue, however, that the only judgement that conceivably comes close is 'would be a disaster.' This seems a stretch though. There is certainly no obvious way in which judging *x* to be a disaster is an instance of the judgement type '*x* is dangerous.' The connection becomes even more tenuous if we consider the more precise evaluative type 'is disagreeably dangerous.' It seems, then, that as the formal object is made more precise it will become increasingly difficult to place concrete cognitions under the correct evaluative category in the course of determining the identity of an emotion.

Of course, S's judgements do seem to have something in common. At the most general level they all conceive of the failed proposal as 'bad.' As a formal object, however, 'bad' is hopelessly vague. It simply fails to differentiate between emotions that we take to be importantly different and distinct. My judgement that my wife's affair has ruined my life is of the type '*x* is bad,' as is my judgement that I am about to be bitten by an angry dog. We would expect, however, each of these occasions to be marked by distinctly different emotions. Similarly, Lyons' quick characterisations of the formal objects of love and grief are unworkably vague. I think any number of things appealing yet do not love them; I believe the Holocaust was a grave misfortune yet I do not, in any

normal sense, grieve over its occurrence.[61] Generally, then, as defining evaluative categories become more general, and hence more inclusive, they lose their bite and fail to mark the difference between importantly distinct emotions.[62] Conversely, as they become more specific in order to more sharply differentiate related emotions they become overly exclusive.

Given the body of considerations discussed above I conclude that the notion of a formal object is insufficient to its fundamental task of accounting for the differentiation of emotions. Of course, the central idea that it embodies—that an emotion's eliciting cognitions in some way determine that emotion's identity—undoubtedly remains compelling. The fact that philosophical accounts of this defining relationship fail under analysis to supply any rigorous principles of identification between emotion and cognition types does not necessarily imply that no principled account can be given. Rather, I would argue that this merely shows the traditional philosophical conceptualisation of this relationship is overly simplistic. It has been hampered, I would suggest, by the two related features I have so far identified as definitive of the hypercognitivist approach to emotion, namely, a dependence upon conceptual analysis to

---

[61] It is worth noting that despite the centrality of the notion to his conception of emotion, with the exception of offhandedly defining fear's formal object as the evaluative category 'is dangerous' (p. 122), de Sousa nowhere ventures to characterise the formal objects of particular emotions. Considered in their particulars formal objects are exceedingly slippery animals.

[62]This point speaks to a counter-objection to my general critique raised by Ronald de Sousa (in conversation): Consider the category 'red' as analogous to formal objects, and the shades of red along the spectrum as analogous to the tokened judgements we are concerned to place under specific formal objects. We have no difficulty in this case placing particular shades under the category 'red', and we can do this even without recognising any single feature shared by all particular shades of red. Why then can we not do the same with tokened judgements and formal objects? Of course, except for problematic fringe instances, we can often do this. We can justifiably place groups of tokened judgements under a single category. My point here is just that this category will most likely be of a level of generality that renders it incapable of differentiating between distinct emotion types.

the exclusion of empirical study, and a misguided focus on the logical relations between cognitive and emotional states.

Cognitive Variance

In addition to these errors there is a third feature characteristic of hyper-cognitivism that has contributed to its misguided picture of the emotion-cognition relationship: a lack of attention to the degree and significance of the variation in the cognitive elements involved in emotion. This indifference stems largely from an uncritical acceptance of the folk psychological taxonomy of cognitive states.

Until this point I have been uncritical in my terminology regarding the nature of the actual eliciting cognition, the classification of which is claimed to be definitive of particular emotions. I have spoken loosely of beliefs, judgements, appraisals, descriptions, and thoughts. This looseness, however, reflects a real variance in the types of cognitive states typically claimed across theories to be involved in emotions. Gordon, for example, argues that belief, a sophisticated cognitive state, is necessary for certain emotions. And while I have sometimes spoken of Aristotle's conception of emotion's defining 'judgements,' I noted earlier that in the *Rhetoric* he actually tends to vary in his description of emotion's cognitive element, shifting between 'appearance' words such as *phantasia* and cognitively stronger and more complex words such as *dokein* and *oiesthai*. Lyons, in turn, discusses the relation between emotions and 'evaluations,' while Robert Solomon identifies emotions with 'judgements.'

Of course, this variety is not in itself problematic. Emotions, as de Sousa notes, *are* variously thought dependent. Some emotions, such as embarrassment, anger, and shame, seem necessarily to involve sophisticated, conscious beliefs.[63] Other emotions seem to arise from significantly different and simpler cognitive states. Merely imagining the death of a loved one might move me to sadness even though I do not believe her to have died. We should thus expect this real variance to be reflected in the terms of any adequate theory of emotion. Hyper-cognitive theories of emotion, however, generally ignore this variance, and fail to take into account important differences in the cognitive components of emotion. These differences are instead usually lumped together under the general concepts noted above: belief, judgement, and evaluation. In this section, then, I want to give some idea of how varied the cognitive foundations of emotion can actually be, and subsequently, discuss some of the particular problems this variance causes for hyper-cognitivism. Perhaps the best place to start is with those emotions like anger and shame that seem to many to most obviously involve uncomplicated belief.

De Sousa's claim that shame, along with embarrassment, being pleased at, and grief, are "founded entirely on belief" stems from his observation that "to change them, all one need do is change the relative belief."[64] He observes: "For my embarrassment to vanish, it is sufficient that I should find out either that no one was watching my faux-pas or that I did not in fact commit one. Grief can be stopped with a word."[65] While undoubtedly true, so far as it goes, the fact that an emotion can be stopped or altered with

---

[63] For de Sousa's claim that anger—along with grief, pity, and compassion—is "clearly grounded in belief," see de Sousa 1990, p. 7. For his claim that shame is similarly grounded, see p. 137. Anger and shame are for Robert Gordon both examples of factive emotions and hence to be angry that *p* or ashamed that *p* a subject must *know* that *p*. See e.g., Gordon 1987, p. 43.

[64] de Sousa 1990, p. 137.

a change in relevant belief does not necessarily prove that that emotion was caused by that or any other belief. This should be unsurprising. We don't as a rule automatically assume that whatever can alter or end some state of affairs must necessarily have caused it; water, after all, extinguishes fire. Merely as a point of logic we thus need not accept de Sousa's claim. More significantly though, a close study of particular cases shows that putatively belief-based emotions like shame can rest on a complex intentional structure that does *not* include belief.

In an article exploring gender differences in the meaning and experience of shame, Sandra Bartky recounts her experience teaching an upper-level extension course to a class of high-school teachers. Her students were an unremarkable mixed group of mature, well-educated professionals. As the class proceeded, though, Bartky noticed that the women in the class tended toward a form of behaviour strikingly distinct from that of the men:

> Though women were in the majority, they were noticeably quieter in class discussion than the men....Women who did enter discussion spoke what linguists call "women's language"....Their speech was marked by hesitations and false starts; they tended to introduce their comments with self-denigrating expressions....In addition to their style of speech, I was struck by the way many female students behaved as they handed me their papers. They would offer heartfelt apologies and copious expressions of regret for the poor quality of their work....Typically [the women] would deliver the apology with head bowed, chest hollowed, and shoulders hunched slightly forward....It became clear to me that many women students were ashamed of their written work and ashamed to express their ideas in a straightforward and open manner. Indeed, it would not be unusual for a student just to say, "I'm really ashamed of this paper," while handing it to me. I have no doubt that these utterances were accurate reports of feeling.[66]

---

[65] Ibid.

[66] Bartky 1990, pp. 88-9

If we accept Bartky's recognition of her female students' behaviour as evidencing some form of shame, and if shame is founded entirely on belief, as de Sousa and others claim, than we should expect the women in Bartky's class to possess the relevant beliefs. What might these be? Analyses differ but most suggest that shame involves the general belief that one has deviated from some public norm or standard and has suffered a resultant lack of standing. This is, of course, just a general schema, like Aristotle's account of anger, which would be filled in by particular beliefs.

Bartky, however, doubts that her female students possess the relevant beliefs:

> I do not think that my students held any such general beliefs about themselves at all; indeed, I suspect that if confronted with such a claim, they would angrily deny it. Could they not point to evidence of past academic accomplishment?...My students felt inadequate without really believing themselves to be inadequate in the salient respects. They sensed something inferior about themselves without believing themselves to be generally inferior at all.[67]

Of course, it is possible that Bartky is simply wrong here; perhaps some or all of her female students *do* explicitly believe that they are inferior in some important way, and it is belief that in fact grounds their shame. What is important here, however, is that Bartky's claim is at least logically possible. There is no good reason to doubt it, other than the circular one that if her female students are ashamed they *must* have the relevant beliefs. Moreover, Bartky's claim is strengthened by her alternative account of the cognitive grounds for the gendered shame she has encountered.

Bartky argues persuasively that the shame of her female students was not the product of explicitly inculcated beliefs about their inferiority, but rather the result of a long immersion in classroom climates that promoted in women a more general

---

[67] Ibid., p. 93.

diminished sense of self. She cites a detailed study on the status of women in education: females are called on less frequently than males; teachers remember the names of male students more often than female students and call upon men by name more frequently; women are praised less than men for work of equal quality; women are interrupted more frequently than men by teachers and other students; teachers make more eye contact with men than with women; in lab courses instructors position themselves closer to male students than to females, and give those males more detailed instruction. The complete list is lengthy.

The end result of these practices, Bartky argues, "is not so much a belief as a *feeling* of inferiority or a *sense* of inadequacy."[68] These vaguer notions of *feeling* and *sensing* are for Bartky distinguished from belief by their being relatively inarticulate and unformed: "...the "feelings" and "sensings" that go to make up the women's shame I describe, do not reach a state of clarity we can dignify as belief."[69] In fact, it is precisely this lack of clarity that is particularly corrosive to women's emotional well being in the classroom. The feelings and sensings that act as the cognitive ground for gendered shame work only by remaining vague and inarticulate:

> Once elevated to the relative lucidity of propositional belief, the suspicion that one's papers are poor, one's remarks stupid, indeed, that one's entire academic performance is substandard, would quickly vanish, overwhelmed by a mass of contrary evidence. With the collapse of these suspicions-cum-beliefs, the shame of which they are said to be constitutive, having no longer any foundation, would just disappear as well.[70]

---

[68] Ibid., p. 94.
[69] Ibid., p. 95.
[70] Ibid.

Bartky thus agrees with de Sousa that change in relevant belief—here the 'change' is actually an elevation of a vague 'sensing' *into* a belief— can sometimes alter or end an emotion. On Bartky's analysis, however, the alteration of an emotion by belief does not preclude the possibility that that emotion was grounded in a complex intentional structure which did not originally include belief.

Amélie Rorty makes a similar point in her discussion of the irrational conservation of emotions. Like Bartky and de Sousa, she allows that "sometimes our emotions change straightaway when we learn that what we believed is not true."[71] Sometimes, however, our emotions are irrationally conserved in that they fail to change appropriately with changes in apparently relevant belief. As an illustration, Rorty considers the case of Jonah, a newswriter, whose anger toward his female boss Esther persists through a long series of changes in belief that should normally have ameliorated his hard feelings. For example, he initially finds her assignments arbitrary and demeaning, but eventually comes to believe he was mistaken in these judgments. Yet his anger towards her continues. He now instead sees her as a petty tyrant. As he continues to work with her though, he comes to grudgingly accept that she is not dictatorial, but is instead quite fair and genuinely interested in her staff's input. Again, however, he remains hostile towards her; every new assignment continues to anger him. Similar changes in belief and understanding continue, but each fails to affect Jonah's anger toward his boss in the way that we would expect *had his anger originally been grounded in or caused by such beliefs*. It thus appears that Jonah's anger was *not* caused by any of his beliefs.

Of course, it might be that we have simply missed the actual belief that grounds Jonah's anger throughout the changes in his other beliefs. Perhaps he is of the general opinion that women in superior positions are not to be trusted and it is this belief that sustains his anger. Or perhaps Jonah simply comes to hold conflicting beliefs about Esther, one of which grounds his belief, and he is guilty of irrational self-deception when he claims to be of one mind about her. These are distinct possibilities. The question we are addressing here, however, "is whether the intentional component of an emotion *always* is a belief."[72] The irrational conservation of emotion might sometimes involve self-deception about belief, or a pure conflict of belief, but there is no good non-circular reason for supposing that this must *always* be the case.

So if irrationally conserved emotions are not to be explained by appeals to belief then how are we to account for them?

Rorty begins her explanation by adopting the standard cognitivist account of the intentional framework of an emotion:

> The immediate object of an emotion is characteristically intentional, directed and referring to objects under descriptions that cannot be substituted *salva affectione*. Standardly, the immediate object not only is the focus of the emotion but is also taken by the person as providing its ground or rationale. The immediate target of the emotion is the object extensionally described and identified.[73]

In Rorty's scheme the "immediate object" is much like the formal object proposed by Kenny, Lyons, and de Sousa. It is an "emotion-grounding description of the [emotion's] target."[74] She thus accepts the basic cognitivist claim that to have an object-directed

---

[71] Rorty 1980, p. 103.
[72] Ibid., p. 115.
[73] Ibid., p. 107.
[74] Ibid., p. 107.

emotion a subject must 'view' that object under a particular description. Rorty diverges from the standard cognitivist account, however, in arguing that this description is sometimes constituted by something other than a belief. She suggests that our most recalcitrant emotions are often instead grounded in patterns of focusing and intentional salience: "when an emotion remains intractable or an anomalous intentional set persists, we suspect that the emotion is rooted in *habits of selective attention and interpretation.*"[75]

More specifically, Rorty identifies a range of intentional components capable of constituting an emotion's grounding description. They include:

> (1) beliefs that can be articulated in propositional form, with well-defined truth conditions;
> (2) vague beliefs in sentential form whose truth or satisfaction conditions can be roughly but not fully specified [such as] "It is better to have good friends than to be rich."
> (3) specific patterns of intentional salience that can be formulated as general beliefs (A pattern of focusing on aspects of women's behaviour construed as domineering or hostile rather than as competent or insecure might in principle be treated as a set of predictions about the behaviour of women under specific conditions....);
> (4) intentional sets that cannot be easily formulated as beliefs (A pattern of focusing on the military defensibility of a landscape, rather than on its fertility or aesthetic composition, cannot be so easily formulated as a set of predictions about the benefits of giving priority to military defense over fertility or aesthetic charm....);
> (5) quasi-intentional sets that can, in principle, be fully specified in physical or extensional descriptions (E.g., other things being equal, painful sensations are standardly more salient than pleasurable ones.)[76]

In Jonah's case, it turns out that his anger is sustained by his continued focus on those aspects of Esther's behaviour that could be interpreted as indicating contempt for males in inferior social positions: the tone of her voice when she assigns Jonah a task; the

---

[75] Ibid., p. 108.
[76] Ibid., pp. 112-13.

quality of her smile (or is that a smirk?) when she offers praise; the differences in her demeanor when she addresses her male *superiors*. The intentional cause of Jonah's anger is thus something like (3) or (4), i.e., a particular *pattern of focusing* on Esther's behaviour that *could*, with an ease and degree of precision dependent upon the details of his actions, be formulated as a belief —in some attenuated sense—about Esther's contemptfulness. Of course, one might ask at this point why we don't simply posit *this* belief as being the *cause* of Jonah's anger. Perhaps Jonah even once held this belief *explicitly* but eventually abandonded it (or so we thought). So again, why appeal to complex talk of "intentional patterns" and "quasi-intentional sets" when we have *some* justification—Jonah's 'attentional behaviour'—for attributing to Jonah a *belief* that could cause his anger?

Rorty's reason for rejecting this move is similar to Bartky's. Both hold that for a state to count as a belief it must reach an adequate "state of clarity":

> Often the only evidence that the person retains the abandoned belief is his emotional state. One of the reasons for resisting assimilating all intentional components of emotions to beliefs is the difficulty of stating what the belief is. There is sometimes no non-question-begging way of formulating a proposition *p*, where inserting *p* in the sentence 'S believes that __' would express the fact that the subject was in that state. A person may not only deny having the abandonded belief but (with the exception of the episode in question) consistently act in a way that supports the denial.[77]

In Jonah's case, the *only* grounds we have for attributing to him the 'belief' that Esther is contemptful toward male inferiors—aside from a circular appeal to his anger—is his recalcitrant habit of focusing on particular aspects of her behaviour. This attribution, however, is decidedly 'loose.' What, for example, would we do if Jonah vehemently and

---

[77] Ibid., p. 115.

sincerely denied thinking that Esther is contemptful? Must we judge him to be irrational? The point here is that Jonah's habit of focusing can simply fail to support the attribution of a state sufficiently distinct to qualify as a belief. In such a case we are thus left to search for the cause of Jonah's anger in the other sorts of intentional components suggested by Rorty.

At this point, having considered two emotions which have seemed to some to most clearly rest entirely upon belief, I now want to move down the scale of complexity and briefly discuss some emotions that seem to fall at the low end of the scale of thought complexity. Here we find some of the emotions most commonly thought to pose difficulties for cognitive theories of emotion. Consider, for example, a simple case: Beethoven's Ninth Symphony inevitably moves me to ecstatic heights so profound I shed tears of joy. What judgment or belief is involved in such a swelling of feeling? My listening to Beethoven might very well produce in me numerous beliefs about his triumphal genius and the grandeur of humanity but it is unclear what belief, if any, could have *caused* my joy. This particular case also illustrates the related problem of objectless emotions. Listening to Beethoven, I seem to be joyful *simpliciter*, not 'joyful about__' nor 'joyful that__'. Such objectless emotions—anxiety and depression are more traditional examples—have traditionally posed a problem for cognitive theorists because in lacking an intentional object they do not stand in need of any cognitive state capable of fixing such an object.[78] Continuing in this vein Paul Griffiths points out that emotions

[78] Realising that objectless emotions pose a problem for his theory, Solomon simply takes objectlessness as indicating that a state is a *mood* rather than an emotion. He thus denies that there really are any objectless emotions. See e.g., Solomon 1993, pp. 70-73. Moods, furthermore, often do have objects; they are simply highly general targets like the "whole of the world." Kenny makes a similar move when considering generalised depression as a possible objection to his claim that emotions are *always* object-directed: "But

stemming solely from flights of imagination pose a further, closely related problem for theorists like Gordon and Lyons.[79] In these cases, almost by definition, no relevant belief, evaluation, or judgment, seems to be involved; I imagine a loved one's death and grow sad, yet at no point do I believe her to have died. Numerous other examples of this sort are already covered rather extensively in the literature so I will not expand here. Instead, with these examples in hand, I want now to look at some of the problems that cognitive variance poses for hyper-cognitive theories.

First, those emotions which fall on the 'simple' end of the thought-dependency spectrum pose a particular problem for the formal object hypothesis. In short, it is not clear whether such simple cognitive states are informationally rich enough to bear an interpretation sufficient to justify those states being placed under linguistically formed and individuated categories of evaluation.[80] Consider, for example, my sudden fearful reaction as I round a corner on the forest path and am confronted with a coiled snake. According to Joseph LeDoux my fear in such cases is initiated by a primitive, subcortical-processing system:

> The visual stimulus is first processed in the brain by the thalamus. Part of the thalamus passes crude, almost archetypal, information directly to the amygdala. This quick and dirty transmission allows the brain to start to respond to the possible danger signified by a thin, curved object, which could be a snake, or could be a stick or some other benign object.[81]

---

are there not objectless emotions, such as pointless depression and undirected fears?...There are indeed such emotions....We are often unaccountably depressed, on days when for no reason everything seems black; but pointless depression is not objectless depression, and the objects of depression are the things which seem black" (Kenny 1963, pp. 60-61).

[79] Griffiths 1997, p. 29.

[80] This is just the point that Rorty and Bartky made with regard to more complex emotions like shame and anger by pointing to instances of those emotions apparently caused by intentional components that were not capable of supporting a propositional interpretation.

[81] LeDoux 1996, p. 166. I will say more about this system in the next chapter.

The question here is whether the "quick and dirty transmission" that is the cognitive/informational state responsible for the elicitation of my response can properly be interpreted as being a contentful state falling under the same evaluative category as the conscious judgement that causes S's fear of the unaccepted proposal. For the formal object hypothesis to hold, and thereby do its intended work of showing why both emotions are instances of fear, the conscious judgement and the "quick and dirty transmission" must share at some level of analysis a content that would allow their identification as instances of the same formal evaluative category. I won't argue here that such an identifications is impossible. It does seem unlikely, however, since it would require ascribing a degree of specificity of content to the "quick and dirty transmission" that would likely be unjustifiable.

Second, and more generally, these same emotions which fall on the 'simpler' end of the thought-dependency spectrum pose the more obvious problem that they stand as factual counter-examples to claims for the necessary involvement of cognitive states like belief and judgement in the production and individuation of particular emotions. For example, 'reflex emotions' like the sudden fearful reaction described above are commonly thought problematic for cognitive theories in general since they occur so rapidly that they could not have been caused by beliefs or judgements, at least as these states are commonly understood. Responses to this criticism are varied, but all are unsatisfactory and most tend toward the obscure. William Alston, for example, simply rules such emotions out as borderline cases.[82] George Pitcher, who like Lyons insists on emotion's necessary ties to evaluative beliefs and judgements, allows that reflex emotions

are problematic but explains them by an appeal to behaviour: "If a person's anger is so great that he makes no conscious evaluational judgement or even has no conscious evaluational belief, then...he acts *as if* he made such a judgement or had such a belief."[83] This observation, while undoubtedly true, is only trivially so, as the point here is to *explain* how such emotions and their accompanying behaviour come about when not initiated by judgement or belief.

Lyons' own explanation of reflex emotions is more substantive than Pitcher's but nearly as obscure. He approaches the problem by arguing that evaluation, as he intends the concept, should be given a dispositional analysis:

> An evaluation can be active but not conscious. That I am afraid of Alsatians is true now though I am writing at my desk. If an Alsatian suddenly appeared I might be plunged instantaneously, reflexly, into a state of fear. Some time ago I formed the view that Alsatians are very dangerous. This evaluation has a structural or categorical basis, a physiological or psychological factor, which lies dormant in me such that it can still be said of me that I believe Alsatians to be very dangerous though I am not thinking of Alsatians at this moment. This factor can be activated instantaneously as a reflex to make me physiologically upset and cause appropriate behaviour as well, most likely, but not to cause any conscious mental acts or episodes which could be labelled as 'evaluating'.[84]

The obscurity here lies in Lyon's appeal to the completely unexplained "physiological or psychological factor" which serves as the "structural or categorical basis" for evaluation. Such obscurity, however, is forgivable, since Lyons intends the mysterious factor only as a placeholder. He expects that future empirical research into "the evaluative part of the brain" will yield more concrete insight into the detailed nature of evaluation.[85] As will become apparent below, Lyons was generally correct in this assumption. I would

---

[82] Alston 1967, p. 324.
[83] Pitcher 1965, pp. 334-5.
[84] Lyons 1980, pp. 88-9.

emphasise, though, that the conceptual framework—the theory of formal objects—in which he embeds his notion of evaluation still faces the problems outlined above. And it will face even further difficulties as we gain the more detailed picture of evaluative processes that Lyons had hoped for. But I will say more about this below.

Conclusion

I have offered this survey of cognitive variance in order to draw the significant lesson that there exists a body of emotions that appear not to require the occurrence of beliefs, judgements, or evaluations, at least as these terms are normally understood. Lack of attention to this fact is a defining feature of hyper-cognitive theories of emotion. This theoretical lacuna, I suggest, stems from the hyper-cognitivist's allegiance to conceptual analysis at the expense of empirical research. More seriously, though, the hyper-cognitive allegiance to purely formal analysis has left them lacking a conceptual framework capable of integrating the valid analytical insights of the cognitivist tradition with emerging empirical facts about emotion. I will try to provide such a framework in the last chapter. At this point, though, I want to outline some of these new facts, since they serve to both reinforce the arguments in this chapter, and to point the way to an approach to emotion capable of overcoming the faults of the hyper-cognitive tradition.

---

[85] Ibid., p. 68.

Chapter Three: The Empirical Study of Emotion

Introduction

The intent of this chapter is to establish two important facts about emotion. I first want to show that there exist in the brain anatomically and functionally discrete neural systems that mediate a significant range of emotions. More importantly though, I want to show that these systems operate to a large degree independently of higher systems in the brain that underwrite the classic exemplars of cognition.

This is an important claim because it helps extend the central argument of the last chapter. The essence of that argument was that philosophers seeking insight into the nature of emotion, and more particularly into emotion's cognitive structure, have typically depended too heavily upon conceptual analysis and the crude categories of folk psychology. In support of this claim I offered some conceptual arguments of my own intended to show that emotions often involve forms of cognition that do not easily fit into the analytical framework philosophy has been led to construct. In this chapter my appeal to the emerging empirically-based understanding of emotional systems is intended to supplement this basic point. It is becoming clear, for example, that the production of emotions by these systems involves the manipulation of informational states that are of a unique type. These states never reach the level of consciousness; they are not inferentially integrated with a subject's explicitly held beliefs; nor are they easily

construed as containing 'propositional content.'[1] Such states thus ill-fit an analytical framework that represents the relation between cognition and emotion in terms of logical relations between propositionally individuated beliefs and sharply defined emotion types.

Beyond further highlighting such deficits, however, the empirical insights I report in this chapter provide philosophy a ground upon which to begin constructing a more adequate understanding of emotion. For example, understanding how many of our emotions are grounded in discrete neural systems that possess their own basic cognitive capacities helps dissolve some of the classical philosophical puzzles about emotion. We no longer need find puzzling objectless, reflex, or irrationally conserved emotions. Similarly, by closely investigating the cognitive capacities of these systems we can begin to limn the cognitive structure of their related emotions from the 'ground up,' so to speak. Philosophy thus need not depend on suspect conceptual analysis.

With these promises in hand I turn now to the details.

The Affective Primacy Hypothesis

I begin here with a discussion of Robert Zajonc's well known "affective primacy" hypothesis, a tripartite claim that prefigures three of the most significant trends in current empirical findings on emotion. As originally intended by Wundt—whom Zajonc credited with first explicitly introducing the idea—"affective primacy" referred to the apparent temporal primacy of affect in consciousness. Wundt claimed that the "affective elements"

[1] Following Stich (1978) I take these three features—accessibility to consciousness, inferential integration, and propositional individuation—as minimally necessary conditions that must be met for a state to count as

of any experience always "begin to force themselves energetically into the fixation point of consciousness *before anything is perceived of the ideational elements*."[2] While Zajonc agreed with Wundt's original claim, he broadened the affective primacy hypothesis by shifting the emphasis from consciousness to the asymmetrical nature of the emotion-cognition relationship. Affect, for Zajonc, is primary in two main ways. First, Zajonc claimed that contrary to traditional understanding the production of emotion did not depend on prior cognition. He argued instead that "to arouse affect, objects need to be cognized very little—in fact minimally."[3] Affect is thus primary in the minimal sense that an emotional response to some object can precede cognition of that object. Temporal primacy, however, is only one aspect of Zajonc's hypothesis.

Zajonc also argued that affect is *phylogenetically primary* to cognition: "...affect is clearly primary in phylogeny....before we evolved language and our cognitive capacities, which are so deeply dependent on language, it was the affective system alone upon which the organism relied for its adaptation."[4] This phylogenetically primary "affective system" is importantly distinct from any later evolved systems that underlie our various cognitive capacities; it is "parallel, separate, and partly independent" of these systems.[5] More specifically, Zajonc speculatively claimed that the affective system is distinct in that it is instantiated in a neural system that is *anatomically distinct* from the higher cognitive systems that reside in the evolutionarily younger neocortex. Perhaps more importantly though, the affective system described by Zajonc is distinct in that it

---

a belief.
[2] Wundt, quoted in Zajonc 1980, p. 152.
[3] Zajonc 1980, p. 154.
[4] Ibid., 1980, pp. 169-70.
[5] Ibid., p. 168.

functions in a fundamentally different way than any cognitive system. Zajonc argued that cognitive systems work by encoding and operating on *discriminanda*, the affectively neutral, extensionally characterisable features of a stimuli—e.g., shape and size—that are involved in pardigmatically cognitive processes like discrimination, identification, and categorisation. In contrast, the affective system works by encoding *preferanda*. I'll say more about these states below, but as a first approximation, preferanda are to be understood as abstract, *emotionally significant*, higher-level properties of a stimulus that figure in our basic affective *evaluations* of stimuli: like/dislike, good/bad, and so on. As described by Zajonc they are "quite gross, vague, and global" and as such are likely "insufficient as a basis for most cognitive judgements—judgements even as primitive as recognition."[6]

The affective primacy hypothesis thus breaks down into the three discrete claims of *temporal primacy*, *phylogenetic primacy*, and what might be called *system independence*. These claims are of course related, but it is important to note that they are distinct; evidence supporting one does not necessarily support the others. Phylogenetic primacy, for example, does not necessarily imply system independence. The fact that a system mediating simple affective responses evolved prior to systems underlying more complex cognitive capacities does not imply that the original system has remained intact and independent, capable of functioning without the aid of higher and phylogenetically newer systems. In short, each claim made by Zajonc must stand on its own. With this caution in mind I turn now to the first aspect of the affective primacy hypothesis.

---

[6] Ibid., p. 159. Much of the subsequent debate on this particular aspect of the affective primacy hypothesis revolves around the fact that derivation of higher level preferenda would seem to necessitate *some*

At the beginning of his argument for the temporal primacy of affect, Zajonc notes the prevailing view that "such cold cognitive processes as recognition or categorisation are primary in aesthetic judgements, in attitudes, in impression formation, and in decision making: They come first."[7] Simply put, it seems that before I can like something, think it beautiful, or form some attitude towards it, I must first know what it is. Counting against this common view, Zajonc argued, is the "mere exposure effect," the well established phenomenon in which exposure to a stimulus is by itself sufficient to induce a preference for that stimulus.

Early explanations of the phenomenon attributed the formation of exposure-induced preferences to positive feelings aroused by the conscious *recognition*—a paradigmatic cognitive process—of familiar objects. Zajonc cites Titchener's account that claimed recognition produced a "glow of warmth, a sense of ownership, a feeling of intimacy."[8] This positive affective response to recognition, in turn, supposedly explained the subject's preference for the stimulus. In response to this traditional explanation, Zajonc cites a number of experiments that show the mere exposure effect does *not* depend upon recognition of the stimulus. In these experiments novel stimuli were presented under conditions that precluded their later conscious recognition. In an early experiment performed by Zajonc subjects wore headphones and listened with one ear to random tone sequences, and with the other to a story that they were asked to track by following a printed version. This diverted their attention from the tone sequences such that when the sequences were presented a second time, without interference, recognition

---

manipulation, however primitive, of lower-level discriminanda. See note 16 below.

[7] Ibid., p. 160.

of these sequences occurred only at chance levels. The subjects, however, still showed a marked preference for the tone sequences that had been previously presented. In another experiment conducted by Zajonc and Kunst-Wilson, a series of random polygons were presented for 1 millisecond, a period too brief to allow for conscious recognition. Again, however, when the original polygons were later presented alongside novel polygons, subjects preferred the originals.

Zajonc's original experiments have since been replicated fairly extensively and his findings extended in various ways.[9] Similar results, for example, have been demonstrated in different experimental paradigms, such as "affective priming." In a typical priming experiment a series of emotionally neutral stimuli—often Chinese ideographs— is flashed on a screen as a subject watches. The ideographs are presented for a period sufficient for conscious recognition. Some ideographs, however, are briefly preceded by 'primes'— photographs of emotionally positive or negative stimuli like often frowning and smiling faces—that are presented for only 4 milliseconds before being blocked by an ideograph. This brief presentation ensures that the primes are not consciously recognised. Despite this lack of recognition, however, the primes seem to influence affective judgements about the ideographs. In one experiment subjects were asked to make simple like/dislike judgements about the ideographs: they consistently liked those that had been preceded by positive primes and disliked those preceded by negative primes.[10] In another trial within the same experiment subjects were asked to guess whether a particular ideograph represented something good or something bad.

---

[8] Ibid., p. 160.
[9] The first replication is in Seamon *et al* 1983.

Again, ideographs preceded by a negative prime were more likely to be judged to represent something "bad" than those preceded by a positive prime, and vice versa.[11]

Zajonc argues that results of these sort prove that "affective reactions to a stimulus may be acquired by virtue of experience with that stimulus even if not accompanied by such an elementary cold cognitive process as conscious recognition."[12] Here, I agree with Zajonc; further experiments discussed below will reinforce this claim. It should be noted, however, that the conclusion drawn by Zajonc is quite limited. The fact that preference for a stimulus can be induced by mere exposure alone, without conscious recognition, does not by itself prove the existence of an independent affective system. The independent system claim is stronger than the temporal primacy claim. To prove this stronger claim it must be possible to show a *dissociation* between conscious recognition of the identity of a stimuius and recognition of the emotional significance of that stimulus, since it is this latter function that partially defines a system as affective. Of course, such a dissociation was found in the priming experiments discussed above, where subjects seemed capable of recognising the emotional value of primes even though they were presented below the threshold of conscious recognition. Other evidence to be discussed below shows a similar dissociation. It thus appears that a Zajonc-type independent affective system does exist. Having said this, however, there remain numerous questions about this system. It is unclear, for example, exactly what basic discriminations it is capable of making. Does the same system underlying the mere

---

[10] Murphy and Zajonc 1993, p. 725.
[11] Ibid., p. 729. The combined effects of suboptimal exposure and priming are studied in Murphy *et al* 1995, with similar results.
[12] Zajonc 1980, p.163.

preference and affective priming phenomena play a similar role in our more complex emotions? Might there be more than one affective system? While I will later suggest some answers to these sorts of questions it is important to raise them here as reminders of the limited conclusions we can draw from Zajonc's own experiments.

It is also important to note that the fact that conscious recognition is not a necessary feature of some basic affective reactions does not preclude the possibility that such reactions are still cognitively mediated *in some basic sense*. This was just Lazarus' point: cognition need not always be conscious and deliberate. I noted above, however, that Zajonc agreed with Lazarus on this issue. Moreover, in his original discussion of the affective system Zajonc allowed that *some* form of discrimination must have occurred, "however primitive or minimal."[13] Why then does Zajonc insist that the process is non-cognitive? One reason undoubtedly involves the primitive nature of the non-conscious processing that underlies the mere exposure and affective priming phenomena. Unlike our more complex cognitive systems it is somewhat limited, capable of making only "gross affective discriminations." It seems, for example, incapable of fine differentiation between closely related but different emotional stimuli.[14] Clearly, though, a simple difference in discriminative capacities is not enough to establish a system as non-cognitive. Again, then, why is the affective system non-cognitive?

---

[13] Ibid., p. 160. This is the sort of qualification that Lazarus picked up on in his criticism of Zajonc to show that Zajonc didn't *really* claim affect could be non-cognitive.

[14] Zajonc describes a relevant experiment: "In a forced-choice discrimination paradigm, participants were exposed to a 4msec suboptimal primes of faces expressing Ekman's six basic emotions. Participants were then shown two faces—an image of the actual prime and an incorrect alternative face, or foil— and asked to guess which of the two faces was the suboptimal prime. Participants made forced-choice discriminations between all possible pairs of Ekman's six basic emotions. Only the positive emotion of happiness was differentiated at a level greater than chance from the negative emotions of anger, fear,

The fuller answer here draws on the second aspect of the affective primacy hypothesis. First, as noted above, Zajonc claimed that the affective system *functions* in an importantly differently way than cognitive systems in that it encodes and manipulates a uniquely distinct class of stimulus features: *preferanda*. Preferanda, recall, are "gross, vague, and global" high-level properties of a stimulus that are unique in that while they can serve as a basis for the emotional evaluation of a stimulus they cannot similarly serve as the basic for recognition or other paradigmatic forms of cognitive judgement. An affective system thus differs from a cognitive system in that it processes different 'types' of information.

While more might be said about the admittedly vague preferanda/discriminanda distinction, the notion is precise enough to present an immediate problem. Put simply, *qua* higher-level property it is difficult to see how preferanda could be constructed without *some* form of manipulation of lower-level discriminanda. Manipulation of information, however, is what seems to define a process as cognitive. This bothered early commentators on Zajonc. For example, Seamon *et al* note:

> To say that the preferanda involve the interaction of "some gross object features and internal states of the individual" comes very close to saying that affective reactions are based on associations made to a particular type of discriminable stimulus feature, and the once clear separation of affect and recognition on the basis of preferanda and discriminanda is lost.[15]

The same problem emerges in later debates about the existence of a uniquely affective systems.[16] The debate, however, is in large part a terminological one that revolves yet

---

sadness, and disgust. Participants were unable to differentiate any of these negative emotions from one another" (Murphy *et al* 1995, p. 600).

[15] Seamon *et al* 1983, p. 553.

[16] Parrott and Schulkin make a similar point against LeDoux, who draws a distinction between cognitive and affective computations that parallels the preferanda/discriminanda distinction: "Cognitive

again around the proper definition of cognition. As such, I don't want pursue this particular topic too far. I raise this point only to suggest that despite terminological problems there still remains an important empirical issue at the heart of the debate. Although critics of Zajonc's position generally assume *as a point of logic* that any 'processing' preceding the functioning of the affective system must be cognitive in some minimal sense, this has not been proven in relation to any fixed definition of cognition. In particular, it has not been shown that preferanda *actually are* constructed out of the *same* lower-level features of a stimulus that support processes like *recognition* of that stimulus, and *this* was the heart of Zajonc's original claim. I don't want to say more here though; lack of empirical data and clear, accepted definitions of cognition render debate on this particular issue fruitless.[17] Instead, I want to turn now to Zajonc's second main motivation for claiming the affective system is non-cognitive, namely, its anatomical independence.

---

computations have as their goal the elaboration of stimulus input and the generation of "good" stimulus representations. Cognitive processing thus leads to more cognitive processing. In contrast...emotional computations have as their goal the evaluation of the significance of the stimulus (determination of the relevance of the stimulus for individual welfare)" (LeDoux 1993, p. 62). Like LeDoux and Zajonc, Parrott and Schulkin recognise that "there are indeed differences between the cognition that is part of fear and the cognition that is part of solving algebra problems" (Parrott and Schulkin 1993a, p. 49). To accommodate these differences they draw a distinction between emotional and "non-emotional cognition," the latter phrase being intended to "emphasise the essential role of cognition in both phenomena" (ibid). Parrott and Schulkin note, however, that "...computations that evaluate the significance of a stimulus presuppose computations about the nature of the stimulus itself" (Parrot and Schulkin 1993b, pp. 67-8). The difference between the two sides thus rests in the different significance afforded the fact that an affective system must apparently receive some cognitively mediated input. For LeDoux, "inputs to the emotional system can be cognitive even if the emotional processing functions are non-cognitive" (LeDoux 1993, p. 62). For Parrott and Schulkin this vitiates any important sense in which the affective system is non-cognitive.

[17] But as we shall see below it turns out that there is good reason for supposing that some forms of evaluation involved in the production of emotions actually do register stimulus properties that are importantly different than those involved in processes like recognition. See, for example, my discussion below of the dissociability of the capacities to (1) recognise the identity of faces, and (2) recognise the emotional content of facial expressions.

Beyond merely functioning in a different manner, Zajonc's proposed affective system is distinguished from cognitive systems in that it is apparently physically realised in a neural system that is *phylogenetically prior* to, and *anatomically distinct* from, the neocortical systems that support the classic paradigms of cognition. In his original paper Zajonc tentatively located the affective system in the limbic system, implicating in particular the amygdala and the hypothalamus, structures that were then known to be *somehow* involved in emotion. However, at the time he wrote his original paper (1980), hard evidence about the neural processes involved in emotion was scarce. Zajonc thus noted in his conclusion that "the language of my paper has been stronger than can be justified by the logic of the argument or the weight of the evidence."[18] Since Zajonc wrote this, however, significant advances have been made in understanding the neuroanatomy of emotion, advances that do carry the weight of evidence for his arguments. I will discuss the neurophysiological evidence in later sections.

At this point, though, I want first to relate some results from experimental psychology in the form of two particularly significant experiments. These experiments serve two main functions. First, both have helped to indicate the basic neuroanatomical structure of the non-conscious affective processing system. Secondly, and perhaps more interestingly, both experiments help reinforce the general claim that some basic emotional responses do not require conscious, deliberate cognition. The first experiment, more particularly, moves beyond simple preference formation to show that the acquisition of conditioned fear of a stimuli does not depend upon possessing relevant declarative knowledge about that stimuli. In this regard it proves one of Zajonc's early conjectures

---

[18] Zajonc 1980, p. 172.

correct. Given the picture of emotion that was beginning to emerge from the mere-exposure experiments, Zajonc had surmised that "it is...possible that we can like something or be afraid of it before we know precisely what it is and perhaps even *without* knowing what it is."[19] The second experiment confirms this finding, but goes further in offering indirect proof for the existence of an affective system capable of recognising the emotional significance of a stimulus—and reflecting this knowledge to consciously accessible levels— in absence of recognition of the identity of that stimulus.

Dissociations of Knowing and Feeling

The first experiment to be related here describes the varying abilities of three subjects, each with a unique form of brain damage, to acquire both declarative knowledge about, and conditioned autonomic responses to, visual and auditory stimuli.[20] The first subject, S1, suffered from bilateral destruction of the amygdala, a small almond-shaped region in the forebrain—implicated by Zajonc as a centre in his proposed affective system—that is emerging as a key neural structure in the production of emotional experience. The second subject, S2, suffered from bilateral damage to the hippocampus. The third subject, S3, had bilateral damage to both the amygdala and the hippocampus. Changes in skin conductivity—alterations in the skin's resistance to the passage of

[19] Ibid., p. 154.
[20] Bechara *et al* 1995.

electrical current resulting from the ANS's subtly increasing production in the skin's sweat glands—served as the measure of the autonomic system's response.[21]

The experiment consisted of two conditioning trials with each trial proceeding in three phases. The general procedure was identical in both trials, although each trial used a different unconditioned stimulus. In the initial habituation phase of the first trial, monochrome slides coloured either green, red, yellow, or blue, were presented in random order until each subject's latent skin conductivity response to the slides approached zero. In the secondary conditioning phase twenty of these slides were again randomly presented to the subjects. Of these twenty, however, six were blue and were immediately followed by a brief blast from a loud boat horn. Immediately following this random presentation of slides six more blue slides were presented but were not followed by a horn blast. The blue slides in this secondary phase thus served as the *conditioned* stimuli, the horn blast as the *unconditioned* stimuli. If the subjects had successfully acquired a conditioned response during this phase then they should have pronounced skin conductivity responses upon the presentation of these last six blue slides as a result of having successfully 'learned' the association between these slides and the unpleasant unconditioned stimulus. In the final extinction phase the subjects were repeatedly presented with blue slides not followed by a horn blast until their skin conductivity response returned to near zero. Finally, after completing the conditioning portion of the experiment each subject was tested for their declarative knowledge about the trial by

[21] Altered skin conductivity as a measure of emotional response is most famously and controversially used in 'lie detector' tests. The principle is straightforward. As the ANS increases the production of the skin's sweat glands the added moisture lessens the skin's resistance to the passage of current. Skin conductivity is measured by passing a low grade current through the skin between two electrodes. Skin conductance response thus measures the change in the amount of current the skin conducts between the two electrodes.

being asked questions of the following type: How many distinct colours were presented? How many colours were followed by a sound? Which colours were followed by a sound? The second trial proceeded in identical fashion to the first but used computer-generated tones in place of the coloured slides. The results of both trials were as follows.

First, during the conditioning phase of both trials all three subjects showed pronounced skin conductivity responses when the stimuli were presented simultaneously with the horn blast. This is significant for two reasons. First, it shows that any subsequent defects in the subjects' responses to conditioned stimuli unpaired with the horn blast—the six blue slides in the latter part of the second phase—cannot be explained as a general inability to generate a normal, heightened skin conductivity response to unconditioned stimuli. That is, given that all subjects had an autonomic reaction when the blue slides were paired with simultaneous horn blasts, any subsequent failure to generate a similar response to blue slides not followed by a horn blast—in short, failure to acquire a conditioned response—is not explicable as a general failure to generate autonomic responses. Secondly, this result also indicates that neither the amygdala nor the hippocampus are necessary for the generation of autonomic responses to unconditioned stimuli. I turn now to results for the individual subjects.

In both trials S1 completely failed to acquire any conditioned response to the stimuli. After the first trial 's initial conditioning phase in which the twenty slides were presented, S1's skin conductivity remained unchanged when the subsequent six blue slides were presented. There was a similar lack of response when the computer-generated tones that had been paired with the horn blast were sounded. On this measure then, the damage to S1's amygdala seems to have prevented the acquisition of a conditioned

response.[22] Despite this failure to acquire a conditioned response, however, S1 was able to acquire complete factual or declarative knowledge about which tones and slides had been paired with the sound. S1 *knew*, and was able to verbally report, which colour of slide and which tone had been followed by a horn blast.

In complete opposition to S1, S2 was able to acquire a conditioned response to both the visual and auditory stimuli. When the last six blue slides were flashed in the conditioning phase S2's skin conductivity spiked, indicating that an autonomic response to the stimuli had been initiated. S2 responded similarly to the conditioned tones. Most significantly, though, *S2 was completely unable to acquire any relevant factual knowledge about the experiment.* For example, he could not report which slide or tone had been paired with the horn blast, nor could he say how many different distinct colours and tones had been paired with a sound. S2 thus lacked the ability to acquire declarative knowledge about the stimuli *even while he could generate a conditioned autonomic response to them.* Finally, S3, the subject with bilateral damage to both his hippocampus and amygdala, failed on both accounts. He never acquired a conditioned response to the tones and slides, nor was he ever able to report any factual knowledge about the experiment.

The authors of this experiment argue that these results demonstrate what they call a "double dissociation of conditioning and declarative knowledge."[23] The dissociation here is 'double' in the sense that the connection between conditioning and declarative

---

[22] Of course, some other autonomic response such as increased heartrate might have been initiated but left unmeasured. Further experiments outlined below, however, suggest that this was unlikely as similar results are obtained using different measures of autonomic response.

[23] Bechara et al 1995, pp. 1117-1118.

knowledge can apparently be severed in two ways. First, S1 *knew* that certain colours and tones were significant; she was conscious of, and able to verbally report, which colours and tones were followed by a loud, obnoxious sound. Possession of this declarative knowledge, however, was *not sufficient* for generating a normal conditioned autonomic response to those tones and colours. This deficit points to two conclusions. First, it indicates that an intact and properly functioning amygdala, while unnecessary for generating autonomic responses to unconditioned stimuli, *is* necessary for emotional conditioning. So while the amygdala is not necessary for generating autonomic responses to unconditioned stimuli—recall that all subjects showed responded to the horn blasts that immediately followed the first six blue slides—it is necessary for establishing relationships between stimuli.[24] More interestingly though, S1's conditioning deficit indicates that her declarative knowledge about the colours and tones and their relationship to the unpleasant sound was *not sufficient* to endow those colours and tones with affective significance. Generalising broadly, it thus appears that *consciously knowing that a stimulus is affectively significant is not a sufficient condition for generating a normal autonomic, affective response to that stimulus.*

Conversely, S2 was unable to acquire any knowledge about which colours or tones were significant and yet was still able to generate a normal response to the affectively significant colours and tones. Contrary to the above case this indicates that *consciously knowing that a stimulus is affectively significant is not necessary for*

[24] Ibid., p. 1117.

generating a normal autonomic, affective response to that stimulus.[25] S2 did not need to know or believe that a blue slide would be immediately followed by an unpleasant sound in order to react autonomically to that slide in a way identical to normal control subjects who did know what would follow the slide. This suggests, though it does not prove, that in normal subjects the conditioned response to a stimulus does not rely upon declarative knowledge of that stimulus. In sum, therefore, in at least this specific case relevant knowledge or belief about a stimulus is neither necessary nor sufficient for the production of a basic emotional reaction to that stimulus. The problem that this poses for the sort of theories discussed in the previous chapter should be clear.

The second experiment that I want to discuss involved the presentation of paired slides to a split-brain patient.[26] The experiment proceeded in two courses. In each course the first slide—the target stimulus—was immediately followed by a second slide—the masking stimulus—that completely blocked the first. The target stimuli were either disgusting—photos of rotted food, skin disease and bleeding wounds; sexual—photos of nudes and erotic scenes; or emotionally neutral—photos of landscapes and various common objects. In the first course of presentations the target stimuli were presented for a period below the subject's previously established threshold of conscious perception. In the second course the target slide was presented for a period above this threshold.

During each course the subject's heartrate was continuously monitored as he performed a number of tasks following each presentation of a target-mask pair. In the

---

[25]The results of S2's trials also indicate something about the function of the hippocampus. Whereas the amygdala seems essential for associating stimuli with autonomic responses, the hippocampus seems essential for learning about the relationship between stimuli (ibid., p 1117).

[26] Ladavas *et al* 1993, pp. 95-114. The use of a split-brain patient was intended to prove the presence of lateralization effects, which I will not discuss here.

"emotional recognition" task he was asked to indicate whether the slide that had preceded the mask was emotionally significant or emotionally neutral. In the "stimulus identification" task the subject was asked to verbally identify the content of the target slide. In both of these tasks he was encouraged to guess even when it seemed, as it usually did, that no slide had preceded the masking slide. Further, the subject also performed two forced-choice tasks at the end of each course. In the first, he had to make a categorical judgement about the target-stimulus along a single dimension—was it living or non-living? This was intended to control for the possibility that differences in the emotional recognition and verbal identification tasks could be a function of their involving the processing of different amounts of information. The second forced-choice task involved stimulus recognition. After each target-mask presentation, two more slides were projected on the screen for two seconds. One was the target while the other was a different stimulus that belonged to the same category as the target—disgusting, sexual, or neutral. The subject then had to choose which slide matched the target. By using stimuli of the same category the experiment was able to control for the possibility that any discrimination that might occur was based on emotional information. Also, both forced choice tasks were carried out immediately after the presentation of the target-mask pair. This controlled for possible memory effects, since the emotional recognition and stimulus identification tasks were always performed ten seconds after the presentation of the target-mask pair in order to allow the patients heartrate to normalise. Results of the experiment were as follows.

When asked to judge whether the below-threshold target stimulus preceding the mask was emotional or neutral, the subject answered correctly at a level *significantly*

above chance. In contrast, the subject could not verbally *identify* the target stimuli that had been presented below-threshold. He refused in 30% of the presentations, claiming that no target slide had been presented, and when encouraged to respond nonetheless he generally gave answers that were completely unrelated to the target, reporting "tree" when the target slide had shown a woman, and so on. However, when the target-stimuli were presented above-threshold, verbal identification was at 100%. A similar pattern emerged in the forced-choice tasks. In the below-threshold course of the stimulus recognition task the subject matched the target slide to the correct later slide at only chance levels while the above-threshold targets were matched correctly for all presentations. The categorisation task produced similar results: below-threshold performance was at chance levels, above-threshold performance was perfect. Significantly, though, analysis of the subject's heartrate through all courses showed that in the below-threshold presentations he responded with an increased heart rate to the disgusting and sexual stimuli, but not to the neutral stimuli. A similar response was noted when the target stimuli were presented above threshold.

These results are clearly significant. The subject's capacity to distinguish between emotional and neutral stimuli, even when the stimuli are presented below the threshold of conscious perception, indicates that *recognition of a stimulus' emotional value does not require conscious identification of the stimulus.* Moreover, the increase in heart rate accompanying the below-threshold emotional stimuli indicates that the *non-conscious perception of emotional value can directly affect the autonomic nervous system.* Both results thus echo and reinforce the findings of the last experiment, as well as the results of the various mere-exposure and affective priming experiments discussed

in the previous section. It seems indisputable, therefore, that *there exist a range of basic emotional responses that involve 'cognitive' processes significantly different from those processes that mediate paradigmatic cognitive activities like identification and belief/knowledge formation.*[27]

If we now take this general point as given, as I suggest we must, several questions immediately arise. Most importantly, we want to understand *how* such responses can occur. The common answer here—and the one suggested by Zajonc—is that at least some emotions must be mediated by *independent affective systems*, i.e., systems capable of generating *intelligent* emotional responses without the involvement of systems that mediate higher forms of cognition. Thus the authors of the last experiment conclude: "The results show that the brain has a *specific mechanism* for distinguishing emotional from neutral situations prior to activating the autonomic nervous system, and that the evaluation of the affective significance of the stimuli may occur at different levels not necessarily represented in consciousness."[28] More specifically, the author's suggest that their results are explicable by the subcortical thalamo-amygdala system, identified by Joseph LeDoux, that I mentioned briefly in the previous chapter. Precisely how this affective system works will be discussed at length in the next section.

In addition to understanding how affective systems work, however, we want also to understand their *prevalence*, an issue that can take various forms. We might ask, for example, how many different emotional systems there are. Are each of our common emotion types mediated by its own system? Suppose that we find reason to believe so.

---

[27] This is the same lesson taught to us by the observations of Bartky and Rorty.
[28] Ladavas *et al* 1993, p. 95.

We might then ask how those systems we have identified are involved in the most *robust* experience of their related emotions. Since the emotional responses I have discussed so far have been quite simple, it might be thought that the systems that mediate *those* responses are not capable of mediating more complex versions of those responses without the involvement of higher cognitive functions. Does the system that mediates simply conditioned fears similarly mediate the most robust human experiences of fear and terror? Perhaps these fuller emotional experiences require the more complex forms of cognition that simpler responses do not.

I will address this and related questions in subsequent sections. At this point, though, I want to begin by relating what is currently known about one of the most thoroughly studied emotional systems: the subcortical thalamo-amygdala circuit identified by LeDoux.

The Neurological Foundation of Fear

There is an emerging consensus that a significant range of experiences and behaviours that most would count as instances of fear are mediated by a coherent, evolutionarily primitive, subcortical neural system. Very roughly, this system runs from the sensory thalamus—where initial processing of sensory stimuli occurs—to the central and lateral areas of the amygdala—where basic processing of the affective significance of stimuli occurs—to various nuclei in the hypothalamus that control the individual autonomic responses characteristic of fear.

The first indication of the existence of this system came from studies using direct electrical stimulation of the brain (ESB). It has been known for some time that electrical stimulation in specific brain regions can initiate a range of fearlike behaviours in animals, including freezing and fleeing responses.[29] Early commentators on these experiments argued that these results showed only the existence of discrete motor control systems. They doubted that a more comprehensive fear system—i.e., one responsible for robust conscious feelings of fear—had been discovered since it appeared that animals could not be conditioned to avoid neutral cues that signalled the onset of the freezing/flight producing ESB. This suggested that the ESB had not produced any real, subjectively-felt aversive emotional experience. Later studies using more sensitive measures, however, showed that such conditioning had in fact occurred. More significantly, later ESB studies of the same anatomical system in human subjects showed that, in fact, robust subjective experiences of fear *could* be produced as evidenced by verbal reports. Jaak Panksepp recounts some of these reports: "...one patient said, "Somebody is now chasing me, I am trying to escape from him." To another, onset of stimulation produced "an abrupt feeling of uncertainty just like entering into a long, dark tunnel." Another experienced a sense of being by the sea with "surf coming from all directions."[30]

More detailed proof for the evidence of a discrete fear system has emerged from animal studies using controlled lesioning of neural pathways and chemical tracing technologies. Joseph LeDoux's work is particularly important here.[31] LeDoux was

---

[29] Panksepp 1998, p. 213.
[30] Ibid., p. 214.
[31] While I focus here on LeDoux's seminal work for the sake of simplicity, numerous other researches have done similar work. For a good summary see Davis 1992.

originally concerned with understanding the neurological mechanisms underlying classic instances of fear conditioning. Earlier studies of the limbic system had shown that lesions in this system disrupted the acquisition of conditioned fears but there was no clear understanding of why this was so. In an attempt to explain these findings, and trace more clearly the structure of the fear system suggested by the earlier ESB experiments, LeDoux designed a complex series of lesion experiments that would reveal the flow of information as it passed through the brain in the course of a subject's acquiring and exhibiting conditioned responses. For example, in an early study of auditory conditioning LeDoux began by lesioning the highest station in the chain of auditory processing, the auditory cortex. This had no effect on conditioned responses. He then lesioned a series of progressively lower stations, from the auditory thalamus to lower stations in the midbrain. All of these lesions prevented fear conditioning. This led LeDoux to conclude that fear conditioning requires that auditory stimuli reach the auditory thalamus but not the auditory cortex.[32]

This presented LeDoux with a problem. Subcortical sensory processing structures like the auditory thalamus had traditionally been thought to project only to their higher cortical counterparts that in turn were thought to perform all of the important *cognitive* work on sensory stimuli that underlay the learning of conditioned responses. It was the sensory cortex that supposedly extracted and processed the emotionally significant information about a stimulus, that *discriminated* and *recognised* the stimulus and *associated* it with past *remembered* dangers—all paradigmatically cognitive functions—and subsequently initiated and organised the correct emotional response.

---

[32] LeDoux 1996, p. 152.

LeDoux's lesioning study complicated this traditional picture—as do the studies discussed in the previous section—since it showed that conditioning could occur *without* higher cortical involvement. He was thus led to search for an alternate neural pathway that could explain this possibility.

LeDoux discovered such a pathway through the use of chemical tracing technologies. Injection of a trace substance into the auditory thalamus showed previously unrecognised direct projections to the amygdala, a small structure in the forebrain that had been the focus of earlier electrical stimulation experiments.[33] (Zajonc, recall, earlier implicated the amygdala as a possible centre for his affective system.) Lesioning these thalamo-amygdala projections prevented the acquisition of conditioned responses. This finding was significant because earlier ESB studies had implicated the central nucleus of the amygdala in the *expression* of conditioned autonomic nervous system responses. Lesions of the central nucleus, for example, had been shown to interfere with all classic measures of conditioned fear, including autonomic responses, the release of stress hormones, reflex potentiation, and the behavioural freezing response.[34] The direct thalamo-amygdala connection discovered by LeDoux thus supplemented these findings by showing that information about a stimulus can reach the amygdala—where conditioned responses are initiated and organised—*without* first passing through the higher sensory cortices. Finally, more detailed lesion and tracing studies continued to reveal additional structures within this system. Projections from the sensory thalamus,

---

[33] Ibid., p. 154.

[34] These findings have been accounted for by the discovery of direct projections from the central nucleus to various nuclei in the lower brain which each separately mediated these functions. Severing one of these projections, for example, interferes with one measure of conditioning, but not any other.

for example, do not proceed directly to the amygdala's central nucleus—which controls the autonomic expressors of fear—but instead enter the amygdala at a small area known as the lateral nucleus.[35] This structure is particularly significant since it seems to be the structure within the thalamo-amygdala circuit that is mainly responsible for the basic processing of a stimuli's affective significance.

While there are additional features of this circuit, this rough picture should serve my interests here. The important point is that the basic structure of the circuit is now quite well understood. What I want to focus on now is the emerging picture of the 'cognitive capacity' of this circuit, and in particular, the discriminative capacities of the amygdala, as it is this emerging picture that I believe philosophy should pay attention to.

The most detailed studies of the sensory processing capacities of the amygdala have focused on auditory stimuli in rats and visual stimuli in primates, and have proceeded by measuring the responses of individual neurons to different types of stimuli via surgically implanted electrodes.[36] One study, for example, isolated a group of auditory response neurons in the lateral amygdaloid nucleus of rats, i.e., neurons that responded specifically to sound. Closer study revealed that these neurons reacted especially strongly to tones within the 13-60 kHz frequency range, and also showed that

---

[35] LeDoux 1996, p. 159. The lateral and central nuclei of the amygdala are connected by direct projections, and by indirect projections that pass through two other subregions of the amygdala, the basal and accessory basal.

[36] The logic relied upon in these sorts of experiments is nicely summed up by Dennett: "At low enough levels of afferent activity the question of reference is answered easily enough: an event refers to (or reports on) those stimulus conditions that cause it to occur. Thus the investigators working with fibres in the optic nerves of frogs and cats are able to report that particular neurons serve to report convexity, moving edges, or small, dark, moving objects because these neurons fire normally only if there is such a pattern on the retina" (Dennett 1969, p. 76).

this reaction was not mediated by the auditory cortex.[37] This narrow sensitivity would undoubtedly be mystifying if it was not known that this range matches that of warning calls sounded by rats under threat.[38] Tones of this frequency are thus environmentally significant for rats since, in the normal case, they are reliable cues of danger. We should thus not be surprised to find a class of cells uniquely tuned to a tonal frequency that serves as a 'natural sign' of danger embedded within a neural system that we know functions to initiate and organise appropriate responses to danger. As I have argued above, Descartes taught us this very point.

Other studies have revealed neurons in the rat's amygdala that are sensitised to a slightly more abstract yet similarly environmentally significant feature:

> Some cells [in the lateral nucleus of the rat's amygdala] habituate quickly to repeated unreinforced presentations of the same stimulus, but they respond strongly if the stimulus is changed or if it is paired with an [unconditioned stimulus]. That is, these cells act as *novelty detectors*: they learn to ignore stimuli that produce no consequences, and hence have no emotional meaning.[39]

Relatedly, several studies have revealed a similar particular sensitivity to novelty in neuronal groups in the amygdala of primates.[40] One experiment, for example, isolated a body of roughly four hundred visually responsive neurons in the macaque amygdala that reacted differentially to various 'biologically significant' objects like food and toys. A subset of these neurons, however, responded only to the sight of *unfamiliar* biologically significant objects, even when these objects differed along various dimensions. This

---

[37] While the auditory cortex of the rat does project to the amygdala, cortical involvement in the activation of these particular neurons to the 13-60 kHz tonal range was ruled out on the basis that the amygdala neurons fired *simultaneously* with the neurons in the auditory cortex that were also reacting to the stimulus tone.

[38] Armony and LeDoux 2000, p. 1072.

[39] Ibid., 1072; my italics.

[40] E.g., Ono and Nishijo 2000, p. 1102; Brothers *et al* 1990, p. 202.

selective response thus suggests that it was the *novelty* of the objects—the only common stimulus feature across presentations—that was driving the neural response.[41] Of course it is possible that the neurons' selective response to these particular novel stimuli was a function of some *other* common but unrecognised feature of the different stimuli such as size or rate of onset. Ruling out these different possibilities would require presenting a wider range of disparate stimuli to the neuron while holding constant *only* the novelty of the stimuli. Needless to say, this is a fundamental concern in experiments of this sort.[42] Again, however, we should not be surprised to find amygdaloid neurons particularly sensitised to novelty, given the salience of novelty in the primate environment: sudden, unfamiliar intrusions need necessarily to be recognised quickly. More strongly, though, we might *expect* to find such cells there, given that the amygdala is embedded in the very system responsible for initiating and controlling our instinctive responses to salient, novel stimuli. The thalamo-amygdala system, for example, controls the freeze response and reflex potentiation, both of which serve to help an organism deal adequately with the sudden appearance of salient novelty in their environment. The freezing response is

---

[41] It further appears that when these neurons habituate to a stimulus, i.e., when they stop responding after repeated presentations, the habituation is not the result of simple sensory habituation, but is itself also a function of changes in the stimuli's biological significance (Ono and Nishijo 2000, p. 1103). One particular 'novelty' neuron, for example, which failed to respond to familiar food and non-food objects, reacted strongly to the initial presentation of an unfamiliar human face with a closed mouth. Repeated presentations of the face through 15 trials caused a decrease in the neuron's response. However, when the same face was presented on the 16 trial with an open mouth (a symbol of threat in primates), the neuron's response increased. Another 'novelty' neuron responded strongly throughout a series of presentations of a small unfamiliar bottle. As soon as the monkey took the bottle and bit it, however, the neuronal response decreased. In both cases *the sudden alteration in the neuron's response tracked the change in the biological significance of the stimuli*, in the first case from non-threat to threat, and in the second case from (something like) novel to benign.

[42] See note 47 below for further discussion.

controlled via connections from the amygdala to the central grey; reflex potentiation via connections to the reticulo-pontis caudalis.[43]

A number of studies of the primate amygdala have also shown that it possesses an even wider range of discrete sensitivities to even more abstract classes of stimuli, in particular to various higher level *social* phenomena. There are, for example, neurons in the macaque amygdala specifically tuned to 'recognise' the *approach behaviour* of other primates, this behaviour being an important element in dominance-establishing primate interactions. An early experiment showed that activity in these neurons increased when a human experimenter walked forward toward the macaque, and decreased when the experimenter walked backward away from the macaque.[44] The same pattern of neural activity also occurred when the experimenter approached the macaque by walking backwards and retreated by walking forwards. This ruled out the possibility that these neuron were responding to a particular physical feature of the experimenter. Moreover, the neurons did not respond at all when neutral items, like a familiar roll of tape, were moved toward and away from the macaque. This ruled out the possibility that the original response was to something like 'generalised movement toward the body' as opposed to the specific social approach behaviour. Of course, in this case it is more difficult to establish with specificity that the neurons were responding to the behaviour class 'approach' *in particular*. This experiment fails to rule out, for example, the possibility of a general sensitivity in the amygdala to the category 'large looming

[43] Davis 1992, p. 356
[44] Ono and Nishijo 2000, p. 1105.

object.'[45] Again, however, we should not be surprised to find discrete neural sensitivities to higher level social phenomena like 'approach' in the amygdala, although here the reason is slightly different. While it seems unlikely that the specific thalamo-amygdala system described above mediates by itself the more complex primate responses to approach, it is clear that the amygdala plays *some* central role in such responses, since macaques that have had their amydgalas removed exhibit a particular inability to respond appropriately to the approach of other primates.[46]

In this same vein, several direct measurement studies have also revealed neuron groups in the primate amygdala that respond primarily to faces, and within these groups subsets of neurons have been identified that are particularly sensitised to the *emotional*

---

[45] The methodology of experiments seeking to measure the response of individual neurons to such abstract, higher level stimuli is typically more complex than the example in this paragraph. Because the stimulus presented is usually a visually complicated scene containing numerous elements, experimenters must develop controls which allow them to determine as precisely as possible which feature of the stimulus scene the neuron is responding to. In one ingenious experiment involving the measurement of a single amygdaloid neuron, an extensive range of situations involving macaques was filmed, broken down into discrete stimulus segments—e.g., 'Juveniles', 'Yawns', Hands', 'Rear Ends', and 'Walking'—and transferred to laser disc so that segments could be switched smoothly and quickly as required. As the trial began, groups of different scenes were randomly flashed onto a screen and the macaque's eye movement was monitored so that it was clear which scene she was attending to. Once a response to a particular scene was identified successive control scenes were presented in order to discern more precisely which feature of original scene had activated the neuron: "For example, during the presentation of a screening set, we detected responsiveness to a segment of an animal 'walking' upside down by gripping the fencing material which formed the roof of her enclosure. We then switched to the 'Upside Down' file, which revealed that static upside down views of animals in the pen did not drive the cell. Views of animals seen from other unusual perspectives were also ineffective, as were views of the same individual in other activities. By proceeding in this fashion, through files such as 'Climbing', 'Walking', 'Running', 'Walking-Reverse', and 'Swaying', we were able to identify the stimulus sets which maximally drove the cell, namely locomotion involving alternating limb movements as in walking, trotting, or climbing" (Brothers *et al* 1990, p. 201). Of course, there always remains the possibility that the neuron was responding to some feature of the stimulus that was not controlled for. The certainty with which we assign 'content' to a neuron's response depends upon the range of controls presented. In this particular experiment, for example, a single neuron was isolated that initially responded when the macaque viewed a scene in which two known animals circled one another in an attempt to acquire an orange. A limited number of control scenes were subsequently presented that allowed the ruling out of certain aspects, e.g., "rotation about vertical body axis." Eventually, the experimenters conclude: "This unit could have been responding to: (a) the perception of an animal 'wanting something, (b) any dominance interaction involving taking or keeping, or (c) interactions involving 'approach'...[which is] normally a component of both (a) and (b)" (ibid, pp. 205-7).

content of facial expressions.[47] In one study, for example, a small group of facially sensitive cells in the amygdala reacted especially strongly when presented with faces with open mouths, a significant sign of threat for most primates.[48] Conversely, and more generally, monkeys who have been bilaterally lesioned in the amygdala typically display a marked and persistent inability to perceive threat instantiated in social situations. Monkeys who have had their amygdalas removed, for example, typically fail to display standard submissive gestures to more dominant animals and consistently try to join alien social groups, a highly abnormal form of behaviour.[49]

Finally, while direct measurement studies of the human amygdala have not yet been performed, there is evidence from two different sources that the human amygdala also plays a central role in processing the emotional content of faces. First, while humans with bilateral damage to the amygdala—a rare condition—are capable of *identifying* familiar faces, they are severely impaired in their capacity to recognise the meaning of emotional facial expressions. Ralph Adolphs and Daniel Tramel have studied one such patient, S.M., who suffered near total destruction of her amygdala as the result of a rare neurological disease.[50] While virtually all of her other cognitive capacities and intelligence remained intact, S.M. showed general post-traumatic incapacity to appreciate the meaning and intensity of typical facial expressions of simple emotions like happiness, surprise, and anger. She also showed a particular incapacity to recognise fearful faces.

---

[46] Brothers *et al*, p. 212.
[47] Rolls 1981; Leonard *et al* 1985.
[48] Leonard *et al* 1985, p. 169.
[49] Dicks *et al* 1969; Leonard *et al* 1985. A similar general incapacity to perceive threat has also been noted in rats that have undergone amygdalectomies. Such rats, for example, will play with a cat, crawl on it, and playfully nibble its ear, a completely unnatural behaviour (Davis 1992, p. 361).
[50] Adolphs *et al* 1994.

Despite these difficulties, however, her capacity to identify familiar faces, and to learn to recognise new faces, was completely intact. Such data, Adolphs and Tramel argue, "provide evidence for a double dissociation between processing of facial identity and of facial affect, suggesting that the two are subserved by anatomically separable neural systems."[51] In support of this general claim, functional neuroimaging studies in which MRIs of the brain are taken as a subject performs an assigned task have shown that the amygdala in normal humans is differentially aroused by emotional versus non-emotional faces.[52] Even more specifically, several of these studies have shown, like Adolphs and Tramel's, that the amygdala is particularly sensitive to fearful facial expressions.[53] Such a particular sensitivity further suggests that the "recognition of facial expression might involve distinct processes tuned to specific emotions."[54] This hypothesis is supported by recent studies of subjects stricken with Huntington's Disease who show a specific inability to recognise facial and vocal expressions of disgust.[55]

In sum, then, when considered collectively these experiments and observations constitute an important first step toward tracing the 'processing profile' of the amygdala, and while much work remains, what we know even in this early stage is significant. First, the amygdala seems able to perform a variety of basic discriminative tasks "without the involvement of the higher processing systems of the brain, systems believed

---

[51] Adolphs *et al* 1994, p. 670.
[52] Dolan 2000, p. 1127.
[53] Dolan 2000, p. 1117; Heining *et al* 2000; Iidaka *et al* 2000; Morris *et al* 1996. These studies have all additionally shown lateralization effects; the left amygdala is especially active in the processing of fearful facial expressions.
[54] Halligan 1998.
[55] Halligan 1998; Heining *et al* 2000; Iidaka *et al* 2000; Phillips *et al* 2000.

to be involved in thinking, reasoning, and consciousness."[56] This is significant for several reasons. It first bears out Zajonc's original speculation about the existence of an anatomically distinct affective system that, with respect to its primitive discriminative capacities, parallels the neural systems underlying our more advanced cognitive capacities but is functionally independent of those higher systems. Relatedly, the specific nature of these capacities also lends significant substance to Zajonc's concept of "preferanda." Preferanda, recall, are abstract, emotionally significant, higher-level properties that figure in our basic emotional 'evaluations' of stimuli, but are insufficient as a basis for more classically cognitive judgements like recognition and identification. Zajonc's original motivation for introducing such states was to help explain the specific phenomenon of mere exposure, but the concept clearly seems applicable to the various dissociative experiments discussed in this and previous sections.[57] In each case the amygdala appears to be processing information that is either physically unavailable to the cognitive systems mediating recognition, or is information of a sort insufficient to support recognition. The dissociability of the capacity to recognise the identity of faces from the capacity to recognise the emotional content of facial expressions is a particularly strong example of this. Given these results, then, it seems that preferanda might constitute a real type of unique, functionally significant mental states.

This is an important point for the general argument advanced in this thesis so I want to digress for a moment to expand upon it. In the previous chapter I argued that

---

[56] LeDoux 1996, p. 161

[57] Zajonc actually surmised in his original paper that the recognition of facial identity, and the recognition of the emotional content of facial expressions, were differentially supported by discriminanda and preferanda (1980, p. 159). At that time, however, there was little experimental evidence that the two capacities could be dissociated.

many cognitively oriented philosophical theories of emotion have made a fundamental mistake in relying uncritically on the crude, folk-psychological taxonomy of mental states, a reliance that has led to numerous problems. I would argue now that the sort of empirical work described here helps to solve these problems in that it provides real evidence for a taxonomy of mental states and processes more adequate to the understanding of emotion. Undoubtedly, many emotions do involve beliefs, judgements and evaluations, as these terms are normally understood; undoubtedly, however, many do not. Recognising the existence of states like preferanda and investigating their nature can help us understand these problematic emotions. Of course, we should have expected that the folk-psychological taxonomy would *need* to be expanded once we began to study in detail the actual workings of emotion. The experiments and observations described here are signal examples of what Dennett has called "sub-personal cognitive psychology," i.e., the detailed study of cognitive activities like pattern recognition, stimulus generalisation, and concept learning, that typically occur at levels inaccessible to simple introspection. "It is here," Dennett notes, "that we will find our good theoretical entities, our useful *illata*, and while some of them may well resemble the familiar entities of folk psychology— beliefs, desires, judgements, decisions—many will certainly not (e.g., the sub-doxastic states proposed by Stich)."[58] Simply put, it should not be surprising that our naive psychology lacks the concepts and vocabulary necessary to adequately characterise the information bearing states and processes operative at levels to which we do not normally have access. Cognitive psychology as a whole has recognised this and adjusted accordingly by abandoning any unquestioned allegiance to the categories of folk

[58] Dennett 1987, p. 63.

psychology. Philosophy of emotion, however, has simply fallen behind the curve in this regard.

With this point in mind, I want to return now to consider some of the further implications of our growing understanding of the amygdala. First, the emerging profile of the amygdala's discriminative capacities provides a possible explanation for the various emotional phenomena I have so far discussed. Very generally, in each case subjects exhibited basic emotional responses to various stimuli while lacking relevant, conscious, declarative knowledge or belief about the stimuli. In the last study cited in the previous section, for example, the subject was able to distinguish between affective and non-affective stimuli even though he could not identify them. Similarly, in the first study of the previous section, involving the conditioning of subjects to neutral slides paired with a horn blast, the second subject exhibited a conditioned response to the blue slides even though he never consciously recognised or believed that blue slides signalled the onset of an unpleasant sound. Such dissociations, I have argued, pose a problem for hyper-cognitivist theories that make belief and knowledge out to be either necessary or sufficent conditions for the production of emotion, since in these cases an emotion, while undoubtedly 'simple,' is clearly produced *without* the subject possessing any relevant beliefs or knowledge. By looking at the thalamic-amygdaloid system, however, we can begin to understand how such dissociations are possible. They likely occur when the amygdala 'judges' or 'appraises' a stimulus to be emotionally significant independent of the more complex, conscious judgements of cortically based cognitive systems. Elisbetta Ladavas, the author of the second study, notes:

> If the thalamo-amygdala pathway can process the emotional significance of the stimuli independently of the cortico-amygdala pathway, then it is possible to understand why the patient in this study was able to judge the emotional value of the stimulus without being able to recognise the object....This is because computation of the affective significance of stimuli can be performed by the amygdala on the basis of some stimulus properties prior to and independent of more complex transformations performed by the cortical centres, such as those involved in the recognition of objects.[59]

This basic idea has recently been extensively developed in a number of 'multi-level' theories of the emotion-cognition relationship.[60] The core claim of these theories is that the production of an emotion can involve a range of qualitatively distinct levels of information and that the 'cognitive' processing of this information can similarly take place at qualitatively distinct levels. Howard Leventhal, for example, has developed a theory that recognises three distinct levels of cognitive processing in emotion: the sensory motor, the schematic, and the conceptual.[61] The research cited above certainly bears this general claim out, but current multi-level theories are more expansive, and consider processing 'sites' beyond the amygdala. Without saying more about the details of such theories—though I believe philosophy should pay attention to these details—I want to point out their potential philosophical significance should they prove correct. At the very least, the existence of such discrete levels of cognitive engagement in emotion could help dissolve the long standing philosophical problems of unemotional evaluation and objectless emotions.[62] Very roughly, unemotional evaluations might occur when higher,

---

[59] Ladavas *et al* 1993, pp. 110-11.
[60] The current state of such theories is summarised in Teasdale 1999.
[61] Leventhal and Scherer 1987.
[62] This possibility is also raised by Paul Griffiths. Drawing on the work of Paul Ekman, which I will discuss below, Griffiths proposes that our basic emotions—surprise, fear, anger, disgust, sadness, and joy—are subserved by discrete "affect programs," each of which is driven in part by its own "automatic appraisal mechanism." Based on such a model, it follows that: "Unemotional evaluations occur when a situation is evaluated by higher cognition as having some ecological significance, but the automatic

cortically based cognitions appraise a situation as emotionally significant, but lower emotional systems in the amygdala and elsewhere do not engage. Conversely, objectless emotions might arise when a lower emotional system is engaged independently of the higher cognitive systems. LeDoux, for example, develops a theory of anxiety—the classic philosophical example of an objectless emotion—that takes precisely this approach.[63] Of course, at the current time such claims are somewhat speculative. The degree to which these classic philosophical difficulties can be dissolved by multi-level theories of emotions depends upon the details. It is still unclear, for example, whether *all* of our emotions are subserved by cognitive engagement at different levels; some cognitively complex emotions, such as those highly culture specific emotions that social constructionists focus on, might not involve any lower-level processing of the type I have been discussing.[64] Moreover, as Ladavas point outs, "to say that the system processes the stimulus on the basis of some primitive stimulus feature or crude perceptual features is not enough. We need a model which specifies the functional subprocesses involved in emotional evaluation."[65] While I would argue that the research I have cited in this section actually goes some way toward providing such a model, there clearly remains much work to be done in this regard. At this point, however, I want to close out this section by considering the second main lesson to be drawn from our growing understanding of the amygdala.

---

appraisal mechanism does not recognise this significance. Objectless emotions occur when affect program responses are inappropriately triggered as they sometimes are in epileptics " (Griffiths 1997, p. 98).

[63] LeDoux 1996, pp. 225-266.

[64] See Griffiths 1997, pp. 137-167.

[65] Ladavas *et al* 1993, p. 111.

This second lesson is a methodological one. In short, I want to suggest that the body of work I have discussed here provides a powerful model for a principled, empirically based method for discerning the cognitive structure of particular emotions. As argued in the previous chapter, philosophy has traditionally approached this goal through the use of conceptual analysis and as a result has faced numerous problems. The model I propose here is instead empirically based. It begins with the identification of the neural substrates of a particular emotion. With regard to fear, this first step occurred in the earliest ESB studies of the limbic system and is currently being continued through lesion studies of the amygdala and observations of amygdala-damaged subjects. What such work gives us is a clear picture of both the *structure* and *function* of the neural substrate under investigation. We know, for example, that the thalamo-amygdala system is capable of independently initiating and controlling a range of highly adaptive autonomic and behavioural responses to a variety of environmentally significant stimuli, i.e., those things we should fear.

Having gained such a picture we can move on to the second step in the process, namely, predicting which cognitive dimensions are likely to be operative in the system being studied *given what we know it can do*. We know, for example, that through its control of the freezing response and reflex potentiation, the thalamo-amygdala system helps us deal with a particular range of environmentally significant stimuli—those dangerous to our physical well-being. Additionally, from the study of animals with severely damaged amygdalas, we know that the amygdala is also involved in the recognition of socially instantiated threat. These sorts of facts support various predictions. We might expect, for example, particular sensitivities to conspecific signals

of danger, or to particular factors of environmental dangers. These might be quite general properties, like novelty, or they might be highly specific. For example, we might expect to find particular sensitivities to sinusoidal movement within the fear circuit of animals for whom snakes are natural predators. Such predictions can in turn guide our investigations into the *actual* cognitive capacities possessed by the system being studied. In this stage we seek to build a 'processing profile' of the system under study, and here we can use the sorts of tools discussed above: direct measurement of neuronal activity, functional neuroimaging, stimulus priming paradigms, and so on. We have, for example, directly measured within the primate amygdala a particular sensitivity—or capacity to 'recognise'—the approach behaviour of other primates. In the human amygdala we have observed, both directly through neuroimaging, and indirectly through observation of a brain damaged patient, a particular sensitivity to fearful facial expressions.

Finally, once we have managed to amass a body of relevant data, we can begin to draw principled general conclusions about which cognitions are central to the emotion and emotional system being studied. Initially, this might seem problematic, as we are likely to end up with disparate set of observations. For example, when the data from studies of the amygdala are viewed at their most concrete level we see a rather varied collection of cognitive capacities, both across species and within single species. Across species, for example, we have discovered a mixture of sensitivities to highly particular stimuli like particular tonal frequencies, and to more abstract stimuli like the emotional content of facial expressions. Within the human amygdala, embedded in the fear system, we have discovered sensitivities to facial expressions of fear, as well as a basic capacity to 'learn,' via simple conditioning procedures, that certain stimuli signal negative

consequences. We may, however, begin to unify such disparate data *if we view them at the right level of generality*. This is an important point, because it is here that we find the motivation to ascend to a level of description that could lead to misunderstanding.

At one level of description, for example, the rat amygdala's sensitivity to tones in a discrete frequency range is markedly dissimilar to the human amygdala's particular sensitivity to fearful facial expressions. At a higher level of description, however, one which recognises the *environmental significance* of the different stimuli *for the subject*, the different capacities of the two amygdalas can be seen to serve the same function: both serve to 'recognise' one form of their species' danger signals. In folk terms, both human and rat amygdalas "judge the situation to be dangerous." It would clearly be wrong, however, to ascribe this propositionally individuated content to the causally efficacious informational states produced and manipulated by the different amygdalas. The level of description that we ascend to in the course of unifying disparate data about the cognitive capacities that we discover should thus not be understood as *strictly* reflecting the informational content of the system it refers to. It does not, however, float completely free of that system. The new level of description is instead an example of what Jackson and Pettit call a *program explanation*, i.e., an explanation "that highlights a common feature of a range of cases...but abstracts away from the causally active features of a particular case."[66] Here, the "common feature" that we highlight will be some non-nomic property—e.g., 'dangerous' or 'novel'—of the class of stimuli to which a system is particularly sensitised that explains *why* the *causally efficacious* states of the system—whatever those might be— have the effect that they do, and further explains *why* the

system is constructed as it is in the first place. Why, for example, should the human amygdala be able to independently recognise facial expressions of fear? Because such expressions, in the normal case, signal 'Danger!' We are thus justified in *abstracting away* from the causally efficacious content of the thalamo-amygdala system—in this case something like 'conspecific expression of fear'—to a particular abstract description of that content— 'dangerous'—because this more abstract property captures the significance of that content *for the subject*. We thus adopt this level of description because it allows us to express the more comprehensive understanding of a system's cognitive structure that we gain when we look at the wider *function* of that system's *particular* capacities. I will say more about this in the next chapter.

Of course, as presented here, this methodological model seems highly programmatic. The actual study of emotion is unlikely to proceed so neatly as my outline suggests. We are unlikely, for example, to discover the neural substrate of a particular emotion merely by chance. Such a discovery is likely to be informed by work at one of the higher levels of study. We might, for example, note the complete lack of a particular emotion in an individual who has suffered some form of focal brain damage. This would be likely to then drive us to look within the area of that damage for a discrete neural substrate for the missing emotion. Clearly, however, such cross fertilisation of levels of study is not really a problem for this methodological model, as it is not intended to be a step-by-step prescription of method. Still, there does exist at least one potentially significant problem for the model proposed here. It is unlikely that *every* emotion is subserved by so neat and distinct a neural system as fear is, and if this is the case then the

---

[66] Clark 1989, p. 198.

methodology I have proposed might have limited application. This concern thus returns us to a central question that I posed earlier: how many 'affective systems' might there be?

Fortunately, it turns out that there *is* good empirical evidence to suggest that there are discrete neural substrates for a significant number of important emotions. I will relate some of this evidence in the next section. Before moving on, however, I want to emphasise what I see as the fundamental strength of the model I have proposed here. Simply put, its strength lies in the fact that it is *based* on empirical observation as opposed to conceptual analysis. Conceptual analysis certainly plays a role here, but it is not the foundational one typical in philosophical attempts to discern the cognitive structure of emotion. Its main role instead comes at the end of the process of empirical study.

Emotional Systems

Empirical investigation into the number and nature of discrete emotional systems has typically followed one of two paths. The first tradition studies the brain directly, making use of the sort of technologies displayed in the previous section: ESB, functional neuroimaging, controlled lesioning, chemical tracing, and observation of neurological pathologies. I will discuss the findings of this tradition below. The second tradition, which I want to first consider briefly, has proceeded via the study of facial expressions and behaviour patterns that are exhibited and understood across cultures and through stages of individual development. Paul Ekman is the central modern figure in this approach.

In a series of experiments spanning almost 35 years, Ekman and his colleagues have studied a wide variety of cultures in an attempt to find 'universally' occurring expressions of emotion.[67] In a typical early experiment participants were shown a set of photographs of facial expressions and were asked to choose expressions they recognised. They were then asked to pick from a list of emotion words—"anger," "fear," "sadness," "disgust," "surprise," and "happiness"—the word that they thought best described the emotion shown by each of the recognised expressions. Ekman's first study, in 1966, involving participants from Chile, Argentina, Brazil, the USA, and Japan, found overwhelming agreement in the participant's judgements of the emotive content of individual facial expressions. In each case participants consistently matched the same word to the same photograph. Following this study Ekman and his colleagues performed similar studies in 21 separate countries, and in each case the results were nearly identical: for a limited but significant range of facial expressions, participants consistently agreed upon which emotion was being displayed in a particular expression. These results were subsequently strengthened by studies performed on the isolated, preliterate South Fore culture of New Guinea.[68] In these studies participants were told an emotionally charged story—e.g., about a man losing his child—and were then asked to point to the pictured facial expression that best suited the story. Significantly, the results were consistent with

---

[67] For a comprehensive history of Ekman's various experiments see his afterword to the recent edition of Darwin's *Expressions of the Emotions* (pp. 363-393). This afterword also contains a fascinating history of Ekman's clashes with Margaret Mead and R.L Birdwhistell over the issue of whether universal facial expressions really existed and, if they did, whether they could be given an evolutionary explanation. Ekman reports, for example, Birdwhistell's initial criticism of his early studies: "[Birdwhistell] argued that people had learned their 'universal' expressions from watching John Wayne or Charlie Chaplin on television, not from our common evolutionary heritage" (p. 377).

[68] These experiments were intended to deflect just the sort of criticism levelled by Birdwhistell, i.e., that the participants had all somehow learned their expressions from a common source even though they came

the earlier studies: the illiterate, culturally isolated New Guineans consistently matched the same facial expressions to the same emotions as had the earlier participants from around the world. Finally, in a methodological twist, Ekman performed another study on the South Fore in which he asked them to show what they would look like if they had been the main character in the story. The participants' expressions were then photographed and filmed and these recordings subsequently shown to Americans. The Americans had no trouble judging correctly which emotion was being displayed.

Such results have led Ekman to the conclusion that there exists a set of facial expressions that are both *produced* and *recognised* universally: anger, disgust, sadness, enjoyment, fear, and surprise. While this general claim is now widely accepted, Ekman's work does have its critics; in the current intellectual climate *no* work that proposes universals in human behaviour goes unchallenged, no matter how thorough and well executed the supporting research. Ideological contention aside, however, even amongst those who agree with Ekman's general account there are differing views about the significance of his findings. In particular, it remains unclear exactly what Ekman's work on facial expressions tells us about the *mechanisms* underlying the emotions associated with those expressions.

Ekman himself argues that the universality of facial expressions implies the existence of innate "affect systems."[69] These systems serve to control the complex co-ordination of the disparate elements that constitute the *entirety* of an emotion, not just its facial expression. These elements include the conscious 'felt' quality of emotion;

---

from an array of cultures.

[69] Ekman 1980, pp 81-84.

stereotypical complex behaviours, like fighting, fleeing, and freezing; altered cognitive activity like increased attentiveness and the activation of particular memories; and a range of emotion-specific changes in the ANS. These elements, Ekman notes, typically occur *rapidly* and *automatically*, without awareness or conscious control, and are *organised* in that the different elements are co-ordinated with one another in functional ways. These co-ordinated changes, Ekman further suggests, are set off by an "automatic appraisal mechanism [that] selectively attends to those stimuli (external or internal) which are the occasion for activating the affect program."[70]

Given this conception of affect systems it appears that Ekman's work supports the view that some emotions are subserved by discrete systems that are functionally independent of higher cognitive systems.[71] I would agree with this interpretation, but only to a point. The value of Ekman's work in this regard is limited for two reasons.

First, Ekman's inference to discrete affect systems is almost entirely based on his observations of the nature and distribution of facial expressions. He argues, in short, that since some facial expressions are universal, and since they are complex, co-ordinated, and largely automatic, it follows that the *emotions* they express must be controlled by "some central direction."[72] Whether or not the general form of this inference is valid—from the automatic occurrence of co-ordinated, functional complexes to the existence of a central control mechanism—Ekman's particular inference here is based almost exclusively on data about facial expressions. He has not shown, for example, the concomitant universal occurrence of complex, co-ordinated patterns of ANS activity specific to particular

---

[70] Ibid., p.84.
[71] See e.g., Griffiths 1997, pp. 79-99.

emotions.[73] If we are strict then, Ekman is only allowed an inference to control systems for facial expression. More seriously, though, even if we were to discover concomitant universals in the other components of emotion, Ekman's approach would be unable to answer fundamental questions about the systems that might underlie such universals. This is in part just a complaint about the empirical limitations of Ekman's approach. I would argue, however, that this is a particularly important problem because the validity of Ekman's central inferential move is questionable.

One of the important lessons emerging from cognitive science is that it is possible to have automatic, complex, co-ordinated systems, *without there being any centralised form of control.*[74] While spelling out precisely what this claim means would take us too far afield, the basic idea is straightforward: complex co-ordinated behaviours can often emerge from the *decentralised* interaction of independent local systems. A simple example makes the point. Consider the complex behaviour of a scheduling system that works to co-ordinate processes and processors—e.g., to match jobs to machines—and must continually deal with variations in the size and complexity of those processes, and fluctuations in the fitness and capacity of processors. If we wanted to build such a system how might we proceed? Andy Clark describes two possible solutions:

> A traditional...solution would invoke a centralized approach in which one system would contain a body of knowledge about the configurations of different machines, typical jobs, etc. That system would also frequently gather data from all the machines concerning their current loads, the jobs waiting, and so on. Using all this information and some rules or heuristics, the system would then search for...an efficient assignment of jobs to machines....Now consider the

---

[72] Ekman 1980, p. 82.

[73] While Ekman and others have studied the ANS changes that accompany the universal facial expressions with some interesting results, research on this question is inconclusive. There is as yet no strong proof for universal, emotion-specific ANS changes. See e.g., Ekman *et al* 1983; Levenson *et al* 1990.

[74] For a summary of recent work in this vein, see Clark 1997.

> decentralized solution....Here, each machine controls its own workload. If machine A creates a job, it sends out a "request for bids" to all the other machines. Other machines respond to such a request by giving estimates of the time they would require to complete the job....The originating machine then simply sends the job to the best bidder. This solution is both robust and soft-assembled. If one machine should crash, the system compensates automatically. And no single machine is crucial—scheduling is rather an emergent property of the simple interactions of posting and bidding among whatever machines are currently active. *Nowhere is there a central model of the system's configuration....*[75]

Given the availability of this second decentralized option it would clearly be a mistake to look at such a complex scheduling system and *infer* that it is controlled by some centralized system. Were we to make such an inference in this case we could be flat wrong. The lesson, therefore, is that we should be generally cautious about *any* inferential move from complex coordination to centralized control. Of course, this is not to say that Ekman is wrong. The thalamo-amygdala fear system, after all, actually looks a great deal like an affect system. The point is instead that the inferential move favoured by Ekman is incapable of settling a range of important issues about the mechanisms that produce our emotions.[76] It cannot, for example, prove that our emotions work more like

---

[75] Ibid., pp. 45-46; my italics.

[76] Paul Griffiths considers a similar challenge to Ekman's work in the form of Neil McNaughton's suggestion that our complex emotional responses are the emergent result of discrete "effector systems" that each separately control some individual component of the complex emotional reaction—e.g., ANS changes—in response to particular *system-specific* triggering stimuli (Griffiths pp. 84-86). The separate effector systems would thus *appear* to be centrally controlled in those situations where their individual elicitors *co-occur*. While there is some support for this claim from animal studies that have "factored" normal stimulus situations, Griffiths doubts McNaughtons's account on the grounds that effector systems would only work to produce co-ordinated responses in highly stereotyped situations, i.e., in situations where the unique elicitors of the different effector systems regularly co-occur. Our emotions, however, can be triggered by virtually *any* situation provided that the right sort of judgement is made. It is unlikely, therefore, that *all* the situations that trigger co-ordinated emotional responses will contain the right mix of elicitors. While undoubtedly true, this counter-objection appeals, as Griffiths himself notes, to the fact that emotions are typically triggered by cognitions. Thus if we accept Ekman's inference to centralised control for this reason it must be because at we believe that *cognition itself* must emerge from some 'central' source. It is, however, precisely the point of the 'decentralised' approach to the mind exemplified in the scheduling example that even high-level cognitive activity might be decentralised. Of course, much of the problem here depends largely on what we mean by "centralised" and "decentralised." Both terms, however, are really nothing more than slogans representing the endpoints of a continuum of organisational

the traditional scheduling system than like the soft-assembled system. Addressing such issues necessitates investigation of the neurological details and we cannot reach those through Ekman's work. With this concern in mind, I thus turn now to the neuroscientific evidence for emotional systems.

The neuroscientific literature pertaining to the investigation of emotion has undergone a recent explosion of such size that a complete summary of current findings is beyond the scope of this thesis. The reason for this explosion, however, is that the neuroscientific study of emotions has been enormously successful. It has, in short, revealed that a significant range of emotions are subserved in the brain by a rich and complex array of highly interconnected but relatively discrete neural systems. Some of these, like the thalamo-amygdala system described by LeDoux, are now fairly well understood; others are grasped only in outline. Additionally, the neuroscience of emotions has also begun to reveal the connections between these emotional systems and higher functioning areas of the brains like the prefrontal cortex that subserve some of our most important rational capacities.[77] These particular insights are especially important for philosophy since they serve, as Ronald de Sousa has suggested, to "confirm what philosophers of emotion have been preaching for some time, namely that emotions are an indispensable part of a rational life."[78] While this is an important topic, I won't say much about it in this thesis. I will, however, say something about the relationship between our 'simpler' emotions—those which are clearly subserved by well defined neural systems—

---

possibilities. The lesson again, then, is that that the actual structure of the emotional systems can likely only be known with any accuracy through neuroscience.

[77] See e.g., Damasio 1994.

[78] De Sousa 1996, p. 329.

and our 'higher,' more cognitively complex emotions whose neural foundations are more obscure. At this point, however, I want to look at what some of the most current neuroscience has discovered about those 'simpler' emotions and the systems that mediate them. As a guide I focus here on the account of emotion systems developed by Jaak Panksepp.[79] Again, my intent is not to offer an exhaustive summary of current research; I want only to give a sense of how far the neuroscientific study of emotion has come in pinpointing the neural circuitry that subserves some of our basic emotions.

Very briefly, Panksepp argues that there are at four "blue ribbon" emotional systems in the brain. These include: (1) a motivational '*seeking*' system that promotes a subject's *interest* in their environment, and mediates *anticipatory excitement*; (2) a '*fear*' system that mediates our responses to danger;[80] (3) a '*rage*' system that mediates feelings of anger and aggression, and specific forms of aggressive behaviours; and (4) a '*panic*' system that mediates the feelings and behaviours associated with social separation, especially the separation of young individuals from their parents. In addition to these four basic systems, two of which I will sketch below, Panksepp also identifies three systems that emerge later in an organism's development and mediate their more complex 'social' emotions. These include: (5) a '*lust*' system; (6) a '*care*' system; and (7) a '*play*' system.

Before going on to sketch two of these systems I want to make two brief points. First, we currently have only a vague understanding of the precise relationship between neural circuits of the sort identified by Panksepp and the full-blooded human experience

[79] Panksepp 1998.
[80] The fear system described by Panksepp is essentially the same as LeDoux's thalamo-amygdala system.

of the emotions those circuits 'mediate.' The words used to describe this relationship—words like "mediate" and "subserve"—should thus be understood as placeholders, the meaning of which remain to be filled in as we gain a better understanding of the role these circuits play in the wider economy of the mind. Having said this, however, it *is* clear that these basic emotional circuits are *essential* to our robust conscious emotions. Without an amygdala, for example, the human experience of fear is radically altered and diminished.

From this warning it follows that it would be highly premature to suggest that the traditional folk categories of emotion will neatly reduce to discrete neural circuits or systems. This is not my intent here however. My goal in discussing such circuits is instead to limn their cognitive capacities so that we might better understand their contribution to the cognitive structure of any emotions they might mediate, however the relationship of 'mediation' is ultimately cashed out.

With these caveats in hand, I now turn to the first emotional system identified by Panksepp.

SEEKING.[81] There is now good evidence that a well circumscribed nexus of emotional experiences and behaviours—the *interest* and *anticipatory excitement* that accompanies one's *energetic, investigative engagement* with their environment—is subserved by an anatomically and functionally discrete circuit in the brain. The core of this 'seeking' system is the ventral tegmental area (VTA) of the lateral hypothalamus and the dopamine-specific neural tract that projects through it. Several strands of research have long suggested that this circuit might play some general role in moderating the level

of energetic engagement with one's environment. One of the earliest and most striking indications of this came from the post-encephalitic patients described in Oliver Sacks' *Awakenings*. Sacks' patients had suffered a severe deterioration of the dopamine circuits that left them eerily 'frozen,' suspended in a sort of perpetual physical and emotional disengagement with their surroundings. Despite having spent years in this state, however, applications of L-DOPA, a form of dopamine, revived Sacks' patients in dramatic fashion, allowing them to return to relatively normal forms of interested and engaged involvement in the world.[82]

Somewhat less dramatically, investigators have for some time known that direct electrical stimulation of various points along the lateral hypothalamus of rats elicits a range of enthusiastically performed goal-directed, investigative behaviours. As was the case with the early ESB studies of the fear circuit, these findings originally led researchers to assume that they had merely discovered the 'control' circuits for these discrete behaviours. Through a careful series of manipulations, however, it was gradually recognized that the ESB had actually produced a behaviourally non-specific 'arousal of interest.' Animals that exhibited a particular behaviour when stimulated would quickly switch to another when the goal of the first behaviour was removed, even though the location of the stimulation remained constant. For example, rats that when stimulated displayed vigorous gnawing intended to acquire a piece of food would quickly change to excited licking of a spout when the food was removed and water introduced. A similar

---

[81] The following account of the seeking system is drawn from Panksepp 1980, pp. 144-163.
[82] For a particularly dramatic example of how L-DOPA acted quite specifically upon one patient's generalised indifference and lack of interest in the world, see Sacks' description of the case of Magda B. (Sacks 1973, pp. 67-73).

easy shift between behaviours was observed when the original food was replaced with another type, or when it was just moved to another location. In each cases the rats were just as likely to move to a *new* seeking behaviour, like gnawing or digging, as they were to continue with the old. Such behavioural flexibility thus suggested that the stimulation had produced a behaviourally non-specific drive in the rats that subsequently found a specific expression determined in part by features of their immediate environment.

Further evidence for the existence of a discrete seeking system has emerged from direct neuronal measurement studies. In a number of primate studies, for example, neurons in the lateral hypothalamus have been found that are highly active when the animal is vigorously investigating their surroundings and searching for food, but quickly shut down when the food is found and consumed. Unfortunately, few ESB or neuronal measurement studies of the seeking system have been performed on humans, but some subjects who have undergone direct stimulation of the lateral hypothalamus have reported feeling that "something very interesting and exciting is going on."[83]

Turning now to the 'input' side of the seeking system, it appears that this circuit is activated most directly by lower brain systems that monitor and regulate homeostatic balance. For example, a range of interoceptive neurons that are specially sensitised to particular bodily imbalances—e.g., *thermoreceptors* sensitised to body temperature fluctuations, and *osmoreceptors* sensitised to concentrations of various nutrients in the blood—feed directly into the ventral tegmental area of the hypothalamus, the heart of the seeking system. Such direct connections allow the seeking system to be activated by a

[83] Panksepp 1998, p. 149.

range of vitally important regulatory imbalances, thereby stimulating us to seek warmth when we are cold, and food when we are hungry.

The seeking system is also activated by a range of external "incentive" stimuli, i.e., "stimuli that predict the occurrence of rewards in the environment."[84] Some of these incentive stimuli, such as the sights and smells of a species' typical foods, appear to be 'innate' and unconditional in the sense that they strongly interact with the seeking system *independent of prior learning*. The seeking system, however, can also be activated by neutral cues—stimuli that are not intrinsically associated with any biologically significant feature of a subject's environment—that have somehow come to be associated with environmentally significant rewards. Dopamine specific neurons in the ventral tegmental area of primates, for example, have been shown to respond selectively to such conditioned stimuli. The precise process whereby the seeking system 'learns' to associate neutral cues with reward is unclear, but Panksepp suggests that it involves a mix of cortical and subcortical areas that are known to project directly into the seeking system. Whatever the process, though, there is good reason for thinking that the lateral hypothalamus is the learning 'centre' of this system. Panksepp cites the work of James Olds, who has studied at length the process whereby animals form the knowledge that they are about to be rewarded. Proceeding via the close study of a simple paradigm—a rat learning to anticipate the delivery of food following the sounding of a brief tone—Olds showed that the lateral hypothalamus was among the first neural structures to exhibit changes indicative of learning. These changes were apparent after the first 10 trials, before there were any behavioural signs of learning. As the trials proceeded,

neuronal changes indicative of learning became evident in ever higher brain areas, accompanied by growing behavioural evidence of learning. Changes in the auditory thalamus and its cortical projections were only apparent in the very last trials. The lateral hypothalamus thus seems to 'learn' the significance of the neutral tone *before* the cortex.

It appears, therefore, that like the fear system, the subcortical seeking system is capable of a range of primitive cognitive activity. At the most basic level, it is innately sensitised to a body of particular unconditional external stimuli, including sights and sounds inherently associated with biologically significant rewards. More significantly, though, the seeking system is apparently capable of a basic form of associative learning that allows it to 'recognise' conditioned stimuli that signal the presence of reward.

RAGE.[85] Next to fear, anger and rage are probably the emotions best understood at the neural system level. One of the earliest detailed studies was carried out by Philip Bard, who along with Walter Cannon developed the Cannon-Bard theory of emotion.[86] Seeking to understand the neural foundations of rage, Bard performed a series of progressive lesioning studies on cats that worked from the cortex down to lower regions in the midbrain. After each lesion Bard provoked the cat to test whether the lesion had had an affect on its capacity to express rage. Significantly, the first lesions to the cortex had no effect on the cat's response, a finding that complicated William James' earlier claim that the motor cortex was responsible for mediating the expression of emotion.[87] Progressively lower lesions similarly had little effect, until Bard lesioned the cat's

---

[84] Ibid., p. 156.
[85] The following account of the rage system is drawn from Panksepp 1998, pp. 187-205.
[86] Bard 1929.
[87] While decorticate animals showed normal responses, they did become markedly temperamental, responding strongly to the slightest provocation. This suggested that the cortex played a role in inhibiting

hypothalamus. Lesions in this area caused a significant disruption to their expression of rage. While the cats still exhibited typical angry responses like snarling, hissing, arching of the back, and piloerection, these previously robust and coherently organised responses were now fragmentary and poorly integrated. They also tended to respond to only the most intense and painful stimuli.

These findings led Bard, along with Cannon, to propose that the hypothalamus was an important *general* emotion centre. Along with the thalamus it was thought to form part of a subcortical circuit that mediated a *range* of different emotions. Roughly, the thalamus, acting as a relay for sensory data, was thought to send information about stimuli simultaneously to the cortex and the hypothalamus. The hypothalamus organised effective behavioural and autonomic responses to the stimuli and fed information about these responses back into the cortex. This information, coupled with the cortex's own recognition of the stimuli, was then consciously elaborated by the cortex in the form of emotional feelings. This model thus made sense of the ability of decorticate cats to *express* rage—since bodily expression of emotion was controlled by the hypothalamus and was thus unaffected by ablation of the cortex—while it implied that loss of the cortex removed the ability to consciously *experience* emotion. It was this last point that led Cannon to label the cat's behaviour "sham rage."

While subsequent research has shown the Cannon-Bard theory to be true in spirit, it has also shown it to be mistaken in detail. Although it is now widely accepted that a significant range of human emotions are *subcortically* mediated, it is no longer thought that there is a *single* subcortical emotion centre in the brain, a change that has in fact

---

inappropriate responses.

stemmed from much of the experimental evidence cited in this chapter. Still, Bard's work was important as it provided some of the first proof that cortical involvement was not a necessary feature of all emotions. Moreover, regarding rage in particular, it pointed to the central involvement of the hypothalamus.

Subsequent research has built on this aspect of Bard's work, largely through the use of closely targeted ESB. Based on these studies there is now good evidence for an anatomically and functionally discrete rage system that mediates a range of angry behaviours and feelings. The anatomical core of this system courses from the middle region of the amygdala down through the medial hypothalamus to its terminus in a variety of nuclei within the periaqueductal gray region of the midbrain.[88] Electrical stimulation at any point along this circuit is capable of producing robust behavioural and autonomic symptoms of rage in a variety of species, including humans. Stimulation of this circuit in the human brain also reliably produces a similarly robust subjective *experience* of rage.

Six areas of the brain are known to have strong connections to the rage system. The highest of these include the medial and lateral regions of the frontal cortex. The medial cortex is particularly significant here since it is known to be centrally involved in the computation of forthcoming rewards and is thus thought to play a role in the creation of frustration, a major precipitant of anger. This area of the cortex, for example, contains

---

[88]Panksepp in fact argues that there are *three* anatomically distinct circuits mediating our angry emotions. In addition to the *affective attack* circuit, which I discuss above, Panksepp also proposes a *predatory attack* circuit that underlies aggressive stalking, and an *intermale aggression* circuit that mediates a specifically masculine form of aggression aimed at the establishment of dominance. Various factors point to the independence of these circuits. For example, each circuit is activated by stimulation of different points in the hypothalamus. Additionally, rats undergoing stimulation of the predatory attack system will

neurons especially sensitised to conditioned stimuli that signal positive reward. When such stimuli change, i.e. when a stimulus previously associated with a reward comes to be associated with the absence of that reward, these neurons reverse their firing pattern, thereby tracking the change in the significance of the stimuli. Relatedly, Panksepp also suggests that the rage system is linked at several points with the seeking system, which as we have seen also works to build expectations of reward, thereby creating the conditions for disappointment and frustration. Anatomical evidence for these particular connections is currently weak, but behavioural studies have shown that animals are more likely to bite when stimulation of the seeking system is turned off—a condition normally indicative of the acquisition of a reward—without their being presented with an actual reward. Given the role that frustration has typically been thought to play in anger these particular connections are highly suggestive

Other areas directly connected to the rage system include those centrally involved in (1) the perception of pain and bodily orientation, (2) the monitoring of peripheral autonomic processes like heart rate and blood pressure, and (3) the monitoring of various homeostatic conditions such as hunger and the level of sexual hormones. The rage system, like the seeking system, thus appears to receive a range of inputs. Some involve primitive cognitive activity—the anticipation of reward—while others inform the system of particular objective conditions of the body.

---

voluntarily continue the stimulation through self-administration, while rats undergoing stimulation of the affective attack circuit try to avoid the stimulation.

Conclusion

While there is much more that might be said about emotional systems, the evidence discussed here should be sufficient to make a compelling case for their existence. More importantly, though, the work cited in this chapter provides a basic understanding of the cognitive capacities of these systems. Considered at a general level, the emotion systems described here are capable of independently 'recognising' a range of environmentally significant categories – e.g., dangerous, novel, and pleasurable – without the involvement of systems that mediate higher cognitive functions like recognition and belief fixation. Any theory that seeks to give a full account of the relationship between emotion and cognition must therefore take these capacities into consideration. I have argued, of course, that the hyper-cognitivist tradition in philosophy is ill-equipped to do so. In the following chapter I thus want to sketch a new form of theory that should be more capable in this regard.

Chapter Four: Toward a New Framework

> Cut anything into tiny pieces and it all becomes a mass of confusion.
>
> - Seneca

Introduction: Two Types of Theory

I want to begin this chapter by making explicit an important distinction that has run implicitly throughout much of this thesis, a distinction concisely expressed in Andy Clark's division of cognitive science into two importantly different forms: *descriptive* and *causal*. He describes the two projects as follows.

> *Descriptive cognitive science* attempts to give a formal theory or model of the structure of the abstract domain of thoughts, using the computer program as a tool or medium.
>
> *Causal cognitive science* attempts to give an account of the inner computational causes of the intelligent behaviours that form the basis for the ascription of thoughts.[1]

To grasp this distinction, consider three different ways in which one might construe the cognitive status of a natural language grammar:

(1) If *a* is a competent speaker of a language, *a*'s competence is causally explained by unconscious knowledge of the rules of a grammar for the language. These rules are internally represented by structures in *a*'s head that have the syntax of the natural language sentences describing the rules.

(2) If *a* is a competent speaker of a language, *a*'s competence is causally explained by the fact that *a*'s information-processing capacities are structured in a way suggested by the form of a grammar for the language.

---

[1] Clark 1989, pp. 153-54.

(3) A good grammar for a language is any theory that yields all and only the sentences characterised as grammatical by a competent speaker of the language. Such a grammar need not be unique, nor need it suggest the form or content of any psychologically realistic theory of language production or understanding.[2]

The first option above represents the strongest form of propositional realism. The terms and relations of the model representing the *product* of some domain of thought—here it is well-formed language—are held to be *psychologically real.* Thus, as Dennett explains it, a grammar understood in this way "describes or mirrors real psychological processes occurring in the production or comprehension of sentences."[3] Fodor's language of thought hypothesis is a classic example here insofar as it holds, for example, that for any predicate of a public language, e.g., "bachelor," "there must be a coextensive predicate of the internal language."[4] The second option marks a weaker position that Clark calls "structural psychological realism."[5] Unlike (1), (2) does not require that the rules of the grammar be sententially coded in its psychological instantiation; it requires only that the discrete functions posited by an adequate grammar be isomorphic to the discrete functions of the brain's language producing system. As an example Clark quotes Jerrold Katz: "Componential distinctions between...syntactic, phonological and semantic components must rest on relevant differences between three neural submechanisms of the mechanism which stores the linguistic representation. The rules of each component must have their psychological reality in the input-output operations of the computing

---

[2] Ibid., pp. 154-55.
[3] Dennett 1977, 269.
[4] Fodor 1975, p. 152. See also note 9 below.
[5] Clark 1989, p. 154.

machinery of this mechanism."[6] Finally, (3) is a wholly *descriptive* theory about the cognitive status of a grammar. Such theories do not concern themselves with the psychological reality of the formal models they produce. They instead only require that they be good models of the abstract domain they characterise. Here, Clark quotes Stich: "[A grammar] describes certain language-specific facts: facts about the acceptability of expressions to speakers and facts about an ability or capacity speakers have for judging and classifying expressions as having or lacking grammatical properties and relations....[The grammarian] is building a description of the facts of acceptability and linguistic intuition."[7]

Not surprisingly, these different views on the status of formal models of the domains of thought have led to active debate. The essence of this clash, according to Clark, can be summarised as follows. Psychological realists, upon constructing a descriptively adequate model of some domain of thought, then argue that by inference to the best explanation we should suppose that people *actually think* within that domain by internally representing that model such that the terms and rules of that model—or perhaps just its functional divisions—are isomorphic to real neural structures and processes. Descriptivists, on the other hand, challenge this inference by pointing to various features of the thinker—typically facts of neural physiology and evolutionary design considerations—that make it unlikely that the model under consideration could ever be psychologically implemented. Clark quotes Devitt and Sterelny:

---

[6] Ibid., p. 155.
[7] Ibid., p. 155-6.

> First, we would want evidence that [any grammar] G was a *candidate* for psychological implementation; that the transformational processes it implicated were within the computational ambit of the mind. Second, the very elegance and simplicity of G is rather more evidence *against* than evidence for, it being the grammar our brain is built to use...[since] adaptions are typically *not* maximally efficient engineering solutions to the problems they solve.[8]

Demands of this sort, for example, have underwritten many of the clashes between psychological realists like Fodor and Pylyshyn and those who think that neural structure and evolutionary considerations make it likely that any psychologically real grammar will be implemented in a connectionist architecture.

What I want to suggest now is that the distinctions and clashes noted by Clark provide a useful metric against which we can place much of what philosophy has said about the emotions, and more particularly, what it has said about the relationship between emotion and cognition.

Consider, for example, the theories of emotion developed by Aristotle and Gordon. Both *could* be understood as (loosely) formal *descriptive* models of the relationship between emotion and cognition. Aristotle, for example, offers an extensive analysis of the epistemic attributions that we are licensed to make when a subject is in a particular emotional state. Gordon similarly argues that particular emotion types require particular types of belief and knowledge in the sense that attributions of an emotion 'that *p*' license attributions of the belief that *p*. As previously noted, however, both Aristotle and Gordon understood their descriptive models as *really capturing* something important about the *actual* mechanisms involved in the production of emotion. The definitions

---

[8] Ibid., p. 156; emphasis in the original.

provided by both were, in Gordon's words, intended "to tell us something about ourselves." Of course, neither theorist makes any *explicit* claim about how the models they provide are likely to be implemented in the brain, though I suspect that if pressed they would adopt a position something like (2) above. It is easy to see, however, how both theories could be understood in the stronger sense of (1) if one held a strongly realist position on propositional attitudes as Jerry Fodor does. On a view of this type, to be angry that *p* requires the belief that *p*, and to have that belief is to have tokened in my brain some symbol *p** in my "proprietary inner code" that is isomorphic to the public language token *p*.[9] In this instance the proprietary inner symbol *p** plays a causal role in the production of my anger. This *same* symbol, however, given the right circumstances, could also have caused me to grieve that *p*, to fear that *p*, and so on.[10]

The theory of formal objects, I would argue, could similarly be construed in this strong manner, though I doubt its proponents would do so. The point, however, is that there is a deep ambiguity here regarding the status of the various projects. We are simply

---

[9] "To have a certain propositional attitude is to be in a certain relation to an internal representation. That is, for each of the (typically infinitely many) propositional attitudes that an organism can entertain, there exist an internal representation and a relation such that being in that relation to that representation is nomologically necessary and sufficient for (or nomologically identical to) having the propositional attitude....Attitudes to propositions are, to that extent, 'reduced' to attitudes to formulae, though the formulae are couched in [an isomorphic] proprietary inner code" (Fodor 1975, p. 198).

[10] "To believe that such and such is to have a mental symbol that means such and such tokened in your head in a certain way; it's to have such a token 'in your belief box,' as I'll sometimes say. Correspondingly, to hope that such and such is to have a token of that *same* mental symbol tokened in your head, but in a rather different way; it's to have it tokened 'in your hope box'" (Fodor 1987, p. 17). Fodor intends this passage to express his formulation of one of the central claims of the Representational Theory of Mind: "For any organism *O*, and any attitude *A* toward the proposition *P*, there is a ('computational'/ 'functional') relation *R* and a mental representation *MP* such that [1] *MP* means that *P*, and [2] *O* has *A* iff *O* bears *R* to *MP*" (ibid.). On this picture, therefore, all 'emotions that *p*' must involve the *actual tokening* of the isomorphic inner symbol *p**.

never told where the formal theories of Gordon, Aristotle, and the others, are intended to fit on the spectrum of possibilities represented by (1)-(3).

Wherever they are ultimately intended to fit, however, I would argue that the empirical work cited in the previous chapter, which I take as exemplary of a *causal* approach to the emotion-cognition relationship, suggests that the theory of formal objects, which I take as embodying the hyper-cognitive approach to emotion, is *not* a good candidate for psychological implementation of any form. The main reason, which I have alluded to previously, is that for a range of emotions the informational states involved in their production cannot be adequately characterised by the basic terms employed in the hyper-cognitivist's formal framework. Many emotions, for example, are caused by cognitive states too primitive to qualify as belief. These cases thus ill-fit a formal framework that adopts belief as a fundamental term. I suspect, moreover, that as the empirical study of emotional systems continues this poor fit will only become worse.

Beyond this point, however, as I argued in my discussion of both Gordon's theory and the theory of formal objects, even if we understand the traditional philosophical model of the emotion-cognition relationship as being *purely descriptive*, it still has some significant 'internal' problems. Gordon's claims about the cognitive foundation of fear, for example, lead to the decidedly unintuitive claim that we cannot be afraid of dying; he does not allow us to simultaneously ascribe to a subject the fear of death and the knowledge that she will die. Similarly, the theory of formal objects places unintuitively strong demands on the epistemic components of emotion. Descriptive theories, however, as expressed in (3) above, are intended to limn the abstract domain of some product of thought by making explicit our intuitions about that domain. An adequate descriptive

grammar, for example, should yield those sentences, and only those sentences, that a competent speaker of the language would judge well-formed. The hyper-cognitive approach of Gordon and others, however, fails in this regard. It is thus akin to a formal grammar that has the unhappy result of marking as well-formed a range of sentences that a normal speaker would not judge grammatical.

Despite these problems, however, and despite the persistence of clashes between theorists of the causal and descriptive bent within linguistics and elsewhere, there is no good reason to suppose that one approach is superior to the other, or that a choice must be made between the two. With regard to the clashes of the grammarians, Clark notes:

> (1) The grammars actually being constructed by working linguists are unlikely to be psychologically real. Nonetheless, they are useful descriptions of real properties of natural languages. (2) Theorists whose goal is the construction of models of the brain basis of grammatical competence will need to focus not only on the data and grammars of (1) but also on the structure of the brain, psycholinguistic evidence, and even, perhaps, evolutionary conjectures concerning the origins of speech and language.... In short, *what is needed is clarity concerning the goals of various studies*, not a victory of one choice of study over another.[11]

This lesson, I believe, applies equally to theories of emotion. I thus want now to explore the idea that the descriptive and causal approaches to emotion discussed in the second and third chapters can be usefully merged to create a formal descriptive model of the domain in which emotion and cognition converge, a model that is truer to the neural and evolutionary facts about emotion, and that additionally avoids the internal problems typical of the philosophical descriptive approach. Having already discussed some of

---

[11] Clark 1989, pp. 156-57; my italics.

these neural facts, I want now to briefly discuss some evolutionary considerations, before going on to sketch this new model.

Evolution and Categorisation

One way of understanding the central idea expressed in the theory of formal objects is that the elicitation and individuation of emotions involves an act of *categorisation*. Philosophy has traditionally portrayed this act as consisting of the tokening of a linguistically individuated judgement that the relevant stimulus *S* has some property *p* in virtue of which *S* is *indirectly* judged to be an exemplar of some category *C*. This category *C* is in turn definitionally/logically related to some emotion type *E*, and it is this relationship that lends *E* its unity. Thus in Aristotle's analysis of anger the defining act of categorisation is the tokening of a judgement that, say, the remark you made to me just now was contemptful. Since contempt, according to Aristotle, is a sub-category of the superordinate category 'slight,' my tokened judgement *in effect* categorises your remark as an *exemplar* of slight. In turn, the superordinate category of slight is—according to someone like Kenny—logically related to the emotion type 'anger.' It is this linkage, from tokened sub-categorisation to logically related superordinal category to logically related emotion type that provides for the identity of the emotional state elicited by the original act of judgement.

I have argued against this view, however, on the grounds that tokened category judgements seldom place stimuli under superordinate categories that are clearly related to emotion types. I have suggested that such a neat alignment can fail to occur for several

reasons. For example, the category under which a tokened judgement places some stimulus might not be a subcategory of any well defined and *relevant* superordinate category. When I fear the rejection of my offer of marriage because I judge such a rejection to be disastrous, what superordinate category, if any, have I implicitly placed such a rejection under? On Lyon's analysis fear's defining superordinate category is 'disagreeably dangerous' but it is unclear why we should view 'disastrous' to be a subcategory of 'disagreeably dangerous.' Conversely, if such cases lead us to redefine superordinate categories in a more general way—we fear that which we judge to be 'bad'—the generality of such categories renders them too vague to differentiate between related but importantly different emotion types. Finally, an alignment of categories and emotion type may fail because the actual cognitive processes that elicit a particular emotion can involve states that are too 'primitive' to support interpretations precise enough to allow them to be placed under particular superordinate categories.

Of course, despite such failures it seems that something like the traditional philosophical picture must be right, as it expresses the undeniably compelling intuition that our various emotion types are loosely but still *systematically* related to particular categories under which a subject views the object of her emotion. What I want to argue now is that the work cited in the last chapter, coupled with some evolutionary considerations concerning the relation of emotion systems to categories, explains *why* this intuition is largely correct, and further, suggests a new way of modelling the cognition-emotion relationship.

To begin, John Tooby and Leda Cosmides argue that there is good evolutionary rationale to suppose that strong selective pressures exist toward the creation of

psychological mechanisms of categorisation that structure an organism's perceptual world in ways most useful to adaptive action:

> A system of categorisation that experiences each event in the world as unique is useless for making decisions. Natural selection, therefore, will act on the organism's systems of categorisation, so that each encounter with the world is perceived and processed in terms of instances of recurring categories. What makes a particular partitioning of events into classes useful to the organism is whether a decision rule based on that categorisation leads to adaptive outcomes. For example, deciding between fleeing or not fleeing requires categorising situations by the cue "predator present" / "predator absent."[12]

It follows, of course, that the categories an organism develops will be useful to the degree that they accurately represent those aspects of reality most important to an organism's well-being; some categorisation schemes will be more useful than others. Some categories, moreover, like that of "'predator present" / "predator absent,"' will represent features in an organism's environment that impact especially strongly on that organism's well-being.

This last point is especially relevant here because emotional systems are commonly understood as evolved responses to precisely this sort of class of recurrent, highly significant situation.[13] Paul Griffiths, for example, argues that "affect programs are adaptive responses to events that have a particular ecological significance for the organism. The fear response is adapted to dangers, the disgust response to noxious stimuli, the anger response to challenges, the surprise response to novel stimuli."[14] Cosmides and Tooby similarly argue that emotion systems evolved as responses to

---

[12] Tooby and Cosmides 1990, p. 408.
[13] Descartes was perhaps the first to explicitly formulate this claim.
[14] Griffiths 1997, p. 89.

particular *adaptive problems*, i.e., "evolutionarily long-enduring recurring clusters of conditions that constitute either reproductive opportunities (e.g., the arrival of a potential mate; the reflectant properties of light) or reproductive obstacles (e.g., the speed of a prey animal; the actions of a sexual rival, limited food supplies for relatives)."[15] Jaak Panksepp even builds this claim into the *criteria* that a neural system must meet to *count* as an emotional system: "The underlying circuits are genetically predetermined and designed to respond unconditionally to stimuli arising from major life-challenging circumstances."[16]

It should be emphasized, however, that the various 'situations' mentioned in the above explanations are really *non-nomic properties*—i.e., properties that do not figure in the statement of natural laws—that can be instantiated in an organism's environment in numerous and highly variable ways. An almost infinite number of situations, for example, can count as 'challenging' or 'dangerous' during both the evolutionary course and individual lifetime of an organism, though some will undoubtedly occur more often than others. In fact, it is precisely *because* these properties can be so variously instantiated that there exist selective pressures to develop *cognitive* mechanisms of categorisation capable of tracking them; after all, as Andy Clark points out, "the very idea of cognition has been tied to...the idea of behaviours carried out in the absence of any constant, lawful, and reliable signal from the local environment."[17] Paul Griffiths also makes this point:

---

[15] Cosmides and Tooby 2000, p. 7. See also LeDoux 1996, p.126; and Nesse and Williams 1994, p. 210.
[16] Panksepp 1998, p. 48.

> The local events which possess the properties of being dangerous, noxious, or novel may be very different from one environment to another. If affect programs are to be of significant adaptive advantage to an organism over an evolutionarily significant time period, *it might well have been advantageous for them to be linked to some mechanism which can interpret the broad ecological categories of danger, novelty, and so forth, in the light of local conditions.*[18]

If these claims are correct then we can begin to see *why* the emotional systems mentioned in the last chapter have the cognitive capacities that they do, and more generally, why a given emotion will be linked to a broad category such that any cognition that *in effect* places a stimulus under that category will produce the related emotion. Simply put, there would have been strong selective pressure toward the creation of mechanisms that categorised events in these ways.

I have, of course, made this point previously in regard to the detailed cognitive capacities found in our various emotional systems. I noted, for example, that we should not be surprised to find a sensitivity to conspecific facial expressions of fear embedded within a system responsible for initiating and controlling responses to situations of a sort that warrant fear. There is, however, further good reason for supposing that in many cases these basic cognitive capacities need not be supplanted by more highly advanced mechanisms of categorisation. In many cases a subject need only detect certain easily recognized *cues* that have reliably signaled the presence of those situations.

The idea is a common one. For example, Antonio Damasio suggests that for a range of primary emotions

> we are wired to respond with [such] an emotion, in preorganized fashion, when

---

[17] Clark 1998, p. 374.
[18] Griffiths 1997, p. 89; my italics. I would only add that "mechanism" should be pluralised here.

> certain features of stimuli in the world or in our bodies are perceived, alone or in combination. Examples of such features include size (as in large animals); large span (as in flying eagles); certain sounds (such as growling); certain configurations of body state (as in pain felt during a heart attack). Such features, individually or conjunctively, would be processed and then detected by a component of the brain's limbic system, say the amygdala....Note that in order to cause a body response, one does not even need to "recognize" the bear, or snake, or eagle, as such, or to know what, precisely, is causing the pain. *All that is required is that early sensory cortices detect and categorise the key feature or features of a given entity...and that structures such as the amygdala receive signals concerning their conjunctive presence.*[19]

In Tooby and Cosmides' account, organisms can rely upon a range of cues of different types. Some, for example, are invariant, concomitant features of salient, *perceivable* features of an organism's environment (as when sinusoidal motion indicates the presence of snakes); others are reliable, perceivable indicators of otherwise "nonperceivable but recurrent sets of conditions [as when] the cue "night" predicts the nonperceivable but recurrent condition "situation in which my ability to detect predatory or enemy ambush far enough in advance to take protective measure is very low"."[20] There are still other "relational cues" that depend for their significance upon the context in which they are perceived.

These cues, in turn, are detected by what Tooby and Cosmides call "situation-detecting algorithms," i.e., basic cognitive mechanisms that need only be capable of 'recognising' the presence of such cues—as opposed to having to recognise the *identity* of what the cues signify—and signalling their presence to the systems responsible for initiating an appropriate response. Tooby and Cosmides divide these algorithms into two

---

[19] Damasio 1994, p. 131; my italics.
[20] Tooby and Cosmides 1990, p. 408-9.

types. The first includes "algorithms that monitor for situation-defining cues: These include perceptual mechanisms, proprioceptive mechanisms, and situation-modelling memory."[21] These sorts of algorithms take as inputs "cues that signal the presence of the situation: for example, low blood sugar signals a depleted nutritional state, the looming approach of a large fanged animal signals the presence of a predator."[22] The second type of proposed algorithm—"algorithms that detect situations"—is somewhat more complex. Algorithms of this sort take the output of those algorithms more narrowly sensitised to situation-defining *cues* "and through integration, probabilistic weighing, and other decision criteria, identify situations as absent or present with some probability."[23] These algorithms thus act on the products of the more focused algorithms to produce probability 'judgements' about the identity and significance of the situation *as a whole.*

We have, of course, seen good concrete examples of both types of 'algorithm' in the previous chapter in the form of the particular cognitive capacities exhibited by the different emotion systems. We saw that the seeking system, for example, receives a range of inputs informing it about important homeostatic imbalances, and is also capable of recognizing cues that predict reward. And while Cosmides and Tooby offer no concrete examples of the second sort of algorithm, I would argue that the amygdala's capacity to recognize the abstract situation of socially significant *approach behaviour* fits their description nicely. I would suggest, moreover, that as our understanding of the fine details of emotion increases we will likely find more examples of these sorts of

---

[21] Cosmides and Tooby 2000, p. 13.
[22] Ibid.

capacities, or algorithms. Without speculating further, however, I want now to emphasize an important aspect of these algorithms.

As Tooby and Cosmides point out, the cues fixed upon by the mechanisms of categorisation involved in the production of emotion work *as* cues because they express the *statistical regularities* that constituted the structure of the environment in which an organism has evolved. Or more specifically, they express the statistical regularities that characterized the environment in which the mechanism of categorisation evolved.[24] This is an important point. It was certainly not always the case, in the course of an organism's evolution, that cues of the above sort correctly indicated the presence of that which we *now* take them to. Not every instance of sinusoidal movement is due to the presence of a snake; enemy ambush does not invariably follow nightfall. These and other similar regularities—which following Tooby and Cosmides I will call "invariances"—*did* occur regularly enough, however, to impact upon the design of the adaptive emotional system. This impact is thus *reflected* in the nature of the adaption to that invariance:

> Species are data recording instruments that have directly "observed" the conditions of the past through direct participation in ancestral environments. A specific complex adaption constitutes, in the improbability of its specialization of design, a probability test about ancestral conditions based on an enormous and representative sample of the past. Eyes tell one that light was a part of the [environment of evolutionary adaptedness]. Immune systems tell one that disease was both present and an important selective agent. The presence of psychological mechanisms producing male sexual jealousy tells one female infidelity was part of the human and ring doves EEA.[25]

On this reasoning, the fear system *qua* adaption tells us that 'danger' was part of the

---

[23] Ibid.
[24] Tooby and Cosmides, 1990, p. 388.

human EEA; the rage system tells us that challenge and frustration were also part of that EEA. It is essential, however, to appreciate the nature of these invariances: "...an invariance is *a single descriptive construct*, calculated from the point of view of a selected adaption or design of a given genotype at a given point of time."[26] "Light," "disease," "danger," and so on, are thus *abstract glosses* on the actual, physically instantiated situations faced by an organism in its evolutionary course, situations that we express in our analysis as statistical regularities. Obviously, however, some descriptive constructs will be more precise than others. "Disease," for example, can be decomposed into a set of more specific descriptions of the particular invariances that acted through natural selection to produce the particular, discrete immune functions that collectively constitute the immune 'system.' This decomposition, of course, will be guided by examining the nature of those discrete functions. Similarly, "danger," *qua* invariant category to which the fear system is an adaptive response, can likely be decomposed into a set of more precise descriptions of the invariances that acted through natural selection to produce the particular cognitive categories employed by the fear system.[27] And like the immune system, this decomposition should be guided in part by looking, as I did in the previous chapter, at the discrete functions that collectively constitute the fear system.

What I want to suggest now is that it is at *this* level that we may most profitably seek our most precise descriptive account of the relation between emotion and cognition.

---

[25] Ibid., p. 390.
[26] Ibid., p. 389.
[27] Nesse and Williams (1994, p. 211) make this same comparison: "Just as there are several components of the immune system, each of which protects us against particular kinds of invasions, there are subtypes of

In essence, such an account constitutes a fine-grained gloss on the emotion-relevant environmental invariances that we now see *reflected* in the categories and categorizing mechanisms employed by our emotions. The construction of this descriptive model proceeds in the manner outlined in the previous chapter. I argued there, recall, that once we begin to gain a detailed picture of the *actual* cognitive capacities of our emotional systems, we will be motivated to ascend to a level of description that abstracts away from the particular categories employed by the system under consideration—e.g., "signal of frequency *x*" or "approach of a conspecific"—and unifies those categories under more general headings—e.g., "danger" or "conspecific threat"—that capture the significance of those categories *for the subject*. I said little there, however, about the constraints on the construction of this level of description. Obviously, though, the nature of the categories that we discover actually being used by our emotional systems will be the first constraint. I would suggest now that the evolutionary observations offered here form a second set of constraints. That is, we must consider the significance of the *particular* categories in relation to the *evolutionary development* of the subject.

So what might this form of description that I am proposing look like?

While it is too early to know the details of the final form, the general shape it must take is discernible. Very roughly, this level of description will consist of a *collection of categories* that are (1) abstract glosses on the causally efficacious categories employed by our emotion producing systems, arrived at via the interpretational

---

emotion that protect us against a variety of particular kinds of threats."

constraints mentioned above, that are (2) subsequently tied to headings signifying discrete emotion types. Luckily, there is in fact a model of the emotion-cognition relationship that looks very much like this: appraisal theory. In the following section I will thus explore the possibility that appraisal theory, *properly construed*, can serve as the foundation for the new level of description that I am proposing.

Dimensions of Appraisal

Following in the spirit of the traditional philosophical approach, modern appraisal theories of emotion similarly focus upon the cognitive processes operant in the elicitation of emotion. Magda Arnold's *Emotion and Personality* is typically credited as offering the first sustained example of appraisal theory's defining approach to emotion. Nico Frijda's summary of Arnold's work provides a succinct description:
"Arnold...proposed...that emotions arise when events are appraised as harmful or beneficial, and that different emotions arise because events are appraised in different ways. [Arnold's] book provides a first attempt to describe these different ways of appraisal as *the systematic variation in a small number of appraisal components or dimensions*, thus systematically accounting for the conditions that lead to the different emotions."[28] So expressed, appraisal theory's continuity with philosophy's treatment of emotion is clear. Both approaches recognise the centrality of cognition in the production

---

[28] Frijda 1993, p. 225; my italics.

and differentiation of emotional states. Despite this basic affinity, however, they differ in at least two significant ways.

Methodologically, appraisal theorists do not depend upon the logical analysis of emotion concepts to derive truths about which cognitions lead to particular emotions, unlike the majority of philosophers I have discussed. Appraisal theory instead approaches such questions empirically. Typical methods currently employed include *self-report*—subjects in whom an emotional state has been evoked are asked to report on the judgement process that preceded their experience of the emotion—and *conjecture about imaginary situations*—subjects are read stories in which people experience emotions and are then asked to describe how the people in the story would likely have appraised the situation that caused their emotion. The shortcomings of these particular methods, of course, is that they assume that the contents of the relevant cognitive processes are available to consciousness and can be easily verbalised. As we have seen, however, this is not always the case. As shown in the previous chapter emotions are often partially subserved by subconscious cognitive processes to which subjects have no introspective access. Verbal report and conjecture-based studies might then only be revealing inaccurate, culturally conditioned beliefs and stereotypes about emotions. Appraisal theory must therefore seek supplementary ways to access the cognitive processes it seeks to characterise. I would argue, of course, that the methodologies cited in the previous chapter offer a good start on this project.[29] Despite these methodological

---

[29] For a discussion of the potential contribution of neuroscientific technologies to the delineation of appraisal dimensions, see Scherer 1993a, pp. 16-19.

concerns, however, even where appraisal theorists have restricted themselves to verbal report and conjecture studies, this approach has been partially vindicated in that diverse studies have tended toward a marked convergence upon a single set of appraisal criteria. I will give an account of this emerging consensus below. Now, though, I want to turn to the second feature that differentiates appraisal theory from its philosophical counterpart.

Conceptually, appraisal theory conceives of the cognition-emotion relationship in a somewhat more complex way than philosophy. For most appraisal theorists the operative cognitive element in the elicitation of an emotion is not a *single* judgement or appraisal but is rather a *pattern* of appraisals, each occurring along some fixed dimension. The complete set of these dimensions may be thought of as forming a multi-dimensional space of possible appraisal patterns that exhaustively characterises the possible ways we may 'cognize' a situation such that an emotional response will be invoked.[30] This way of characterising the formal nature of emotion's cognitive element is largely missing from the philosophical approach to emotion, even though it has sometimes been implied. Earlier, for example, I characterised both Aristotle and Descartes as recognising something like a set of dimensions along which our judgements can vary in the course of producing an emotion. This characterisation, however, is only implied in their respective approaches and is never explicitly developed. More recently though, Ronald de Sousa

---

[30] This understanding of the dimensional sets proposed by appraisal theorists is suggested by the similar use of multidimensional scaling made by psychologists who have attempted to find principled ways of characterising environmental scenes which evoke stereotypical forms of behaviour (e.g., see Tversky and Hemenway 1983, pp. 123-124.)

has suggested that we "think of the different formal objects as independent *dimensions of evaluation.*"[31]

While intriguing, these apparent similarities between appraisal theory and traditional philosophical approaches are problematic. For example, de Sousa argues that each emotion type is differentiated by its relationship to a *single* formal object. Thus if we substitute 'dimension of evaluation' for 'formal object,' as de Sousa suggests, this picture would have emotion types individuated by their relation to *single* dimensions of appraisal. This is not what appraisal theory proposes. It instead holds that each emotion type is characterised by judgements along a *number* of dimensions, and that the identity of a particular elicited emotion depends upon the pattern of appraisal across these dimensions. Moreover, as I've argued above, de Sousa's explicit account of formal objects, which follows Kenny and Lyons', characterises them as an intermediary link between tokened judgements and emotion types. The process of analysis that justifies this linkage proceeds via semantic and logical relationships claimed to hold between these three elements. This picture, however, differs fundamentally from that of appraisal theory. I will argue below that appraisal theory's multi-dimensional model of emotion's cognitive element suggests a fundamentally different formalism from that employed by the theory of formal objects, one which allows for a qualitatively different analysis of the relationship between tokened judgements—or particular appraisal patterns—and emotion types. For the moment, though, I want to first give some better idea of the concrete

---

[31] de Sousa 1987, p. 173. Taken collectively, de Sousa argues that these dimensions similarly delimit a range of real features he calls the "axiological."

content of appraisal theories, particularly, their explicit proposals regarding the 'content' of the dimensions that collectively constitute emotion's cognitive framework.

In an exhaustive review of current appraisal theories, Klaus Scherer, a prominent appraisal theorist, identifies eight main appraisal categories, and within those categories, 17 specific dimensions, along which a subject appraises a stimulus in the course of developing a particular emotional response toward that stimulus:[32]

Dimensions of Appraisal

Change of State

Novel.....Expected

Stimulus Event Type

Internal to Agent.....External to Agent

Intrinsic and Extrinsic Hedonic Valence

Intrinsically Pleasant.....Intrinsically Unpleasant
Extrinsically Pleasant.....Extrinsically Unpleasant

Cause of Event

Self / Other / Natural Agent

Nature of Event and Relation to Goals

Goal is Personal / Goal is Relationship / Goal is Social
Effects Self's Goals.....Effects Others' Goals

---

[32] Scherer 1988, pp. 89-126. See also Scherer 1999, pp.638-40.

Impossible.....Unlikely.....Probable.....Certain
Central.....Peripheral

Consistency of Event Consequences with Goals

Helps Achieve Goal.....Blocks Goal Achievement
Consistent With Expected State.....Inconsistent with Expected State
Urgent Response Required.....Response Not Yet Required

Coping Potential

Event Can Be Controlled.....Event Cannot be Controlled
Self Able to Control.....Self Unable to Control
Need to Adjust Goals.....Adjustment of Goals Not Needed

Relation of Event to Moral Standards

Violates Internal Rule.....Accords With Internal Rule
Violates External Rule.....Accords With External Rule

A more comprehensive account of specific proposals could be given, but the picture presented here should be complete enough to draw a clear picture of the dimensions of appraisal typically held to be operative in emotion. With such a picture in hand, however, several questions immediately arise. How does appraisal theory view the actual processes of appraisal? Do they involve conscious deliberation resulting in the production of linguaform judgments? Can appraisals along the same dimension occur at different levels of processing? More basically, is appraisal theory a good theory? Are different, distinct emotion types related by similar patterns of appraisal? Are the dimensions of appraisal upon which theorists appear to be converging the correct ones?

While these are important questions, I won't deal with them here, since they are largely tangential to the way in which I am suggesting we understand appraisal theory. As noted in the previous section, I am suggesting that appraisal theory be read as an essentially *descriptive* theory pitched at a level that abstracts away from the content and nature of the causally efficacious cognitive states involved in the production of emotion, and instead captures the *meaning* or *significance for the subject* of the categories actually processed. The categories cited by Scherer should thus be read as being just those fine-grained glosses mentioned in the previous section that are intended to capture these meanings. On this understanding "appraisal" is thus a general term intended to be neutral in regard to the nature of the cognitive processes under consideration. It does not demand that the processes and states it confounds be conscious or unconscious, nor that they involve propositional or non-propositional content. This is, in fact, a fairly common understanding amongst appraisal theorists of how "appraisal" should be understood. Even while they adopt this view, however, most theorists still see the need to understand the *actual* nature of the causally efficacious states and processes subsumed under the general heading "appraisal." Scherer, for example, explicitly argues for these two related points:

> *...we need a general, overarching term to cover the fundamental fact that it is not the objective nature of a stimulus but the organism's "evaluation" of it that determines the nature of the ensuing emotion.* A completely automatic, reflexive defense reaction of the organism also constitutes an intrinsic assessment, a valuation, of the noxiousness of the stimulus....Even if simple feature detection is involved, the outcome of the process constitutes an assessment of the significance of the detected stimulus to the organisms, given that feature detectors that have any behavioural consequences are automatically "significance detectors." Obviously, this is a different process from the one that allows us to infer that a particular bit of news, given its ramifications, may have negative impact on our plans. And, obviously, the resulting emotional state and the action tendencies

> produced are different. Yet, both types of emotion-antecedent processing share a number of central functional-adaptional aspects. Does one want to emphasize this communality by talking about significance detection as "appraisal" or "evaluation" in the widest sense (without prejudging the nature of the underlying processes)? Or does one want clearly to demarcate the different processes by applying different concepts? The answer depends on the desired level of analysis, on the clarity of the respective definitions, and on the consensus on the conceptual distinctions made. In any case, *researchers in this field need to undertake the in-depth study of the precise mechanisms that are involved.* It would seem that this is where future efforts should be undertaken.[33]

With this understanding of "appraisal" and appraisal dimensions in hand, I want to now examine a recent attempt to evaluate the adequacy of a specific form of appraisal theory, Scherer's expert system GENESE (Geneva Expert System on Emotion). The value of GENESE lies not so much in the degree to which it validates any particular set of appraisal dimensions, even though this was the original intent behind its creation. The relevance of GENESE to my concerns stems instead from the fact that the formalism it employs is particularly well suited to the level of description that I am proposing we pursue, as it allows for a flexible yet strongly principled way of expressing important relationships between classes of cognitions and emotion types that manages to avoid the problems of the propositionally-based descriptive framework favoured by hyper-cognitivism. I turn now to my account of GENESE.

---

[33] Scherer 1999, p. 649; my italics.

GENESE: Toward a New Formalism

GENESE is a computationally implemented expert system founded upon Scherer's own component process model of emotion, a variant of appraisal theory distinguished by its claim that appraisals occur in a fixed sequential order.[34] As implemented in GENESE Scherer's theory proposes fourteen discrete emotion types and for each type hypothesises a prototypical eliciting pattern of stimulus evaluation along fifteen distinct appraisal dimensions.[35] The function of GENESE is to predict which of the 14 emotion labels best describes an actual experienced emotion as described by test subjects. This prediction is based on a comparison of the actual appraisal process that preceded the emotion, as described by the subject who experienced the emotion, to the different hypothesised prototypical appraisal patterns. GENESE returns as its prediction the emotion label that is affixed to the prototypical appraisal process most closely resembling the actual appraisal process.

Unlike typical expert systems, which generally employ series of 'if-then' rules, GENESE employs a multi-dimensional vector space model similar to the state space models used in the formal characterisation of parallel distributed processing or 'connectionist' systems. Each of the fourteen emotion types is related to a prototypical category vector whose value is determined by numerically quantifying the hypothesised

---

[34] Scherer 1984, 1986.

[35] For a summary of the emotion types used in GENESE, and their related prototypical appraisal pattern,

appraisal pattern associated with that type. This quantification is effected by assigning a numerical value to each of the fifteen appraisal dimensions of that pattern. These numeric values translate the 'semantic' values of 'high,' 'low,' 'self,' 'other,' and so on. A particular *prototype vector*—or simply *prototype*—can thus be conceived of as a fixed point in the 15-dimensional vector space. Similarly, an *input vector* for the actual experienced emotion we wish to classify is determined by the subject's answers to fifteen questions, each with predefined numeric answer categories, where each question corresponds to one of the fifteen possible appraisal dimensions. The input vector thus constitutes a second fixed point in the 15-dimensional vector space.

We thus have (1) fourteen fixed prototype vectors that are essentially numeric expressions of the appraisal processes that hypothetically define each of the 14 typical emotion types, and (2) any number of variable input vectors, each one representing the actual appraisal process operant in the production of the subject's emotion we wish to classify. Given these two points GENESE proceeds in this classification by systematically comparing a given input vector to each of the fourteen fixed prototype vectors using Euclidean distance measures. GENESE then returns as the most appropriate classification the emotion label that is associated with the prototype vector that falls closest to the input vector.

Without delving into the actual results of the trials run with GENESE I want to turn now to one immediately problematic aspect of GENESE, namely, the way in which

---

see Appendix B.

it represents both prototypes and their relation to emotion types.[36] Recall: a prototype is a *single* fixed point in the 15-dimensional appraisal space whose location is determined by numerically quantifying a particular appraisal pattern that is hypothesised to differentiate the emotion type designated by the label attached to that vector. This picture presents two problems.

First, it makes it difficult to discern how such a model truly differs from the theory of formal objects, since both approaches seem to claim that emotion types are individuated by their relation to unique descriptions. That is, one way of understanding GENESE's prototype vectors is to view them as encapsulating complex conjunctive descriptions of the typical or 'ideal' object of the emotion designated by the affixed label. Consider, for example, the appraisal pattern used to differentiate fear in GENESE. This pattern gives a description of fear's 'ideal' object: it appears suddenly, is unpredictable, is unpleasant, will likely impact upon our physical body, and so on. On this understanding, however, prototypes differ little from formal objects. Fear's defining formal object, after all, is simply that description which must be apprehended as applying to any particular object that we claim to fear. Thus GENESE appears to differ only in that it posits a more complicated description than the sort typically put forth by philosophers like Kenny, Lyons, and de Sousa.

While this is a valid point, so far as it goes, it is less a problem for appraisal theory than it is an important point of contact between it and the philosophical approach

---

[36] For an account of GENESE's success see Scherer 1993, pp. 341-349.

to emotion embodied by the theory of formal objects. Both theories, in short, agree that what matters in the production of an emotion is the description that the subject 'apprehends' as applying to the object of her emotion. They simply conceive of this act of apprehending in different ways. Formal object theory, and the hyper-cognitive tradition out of which it emerges, holds that it can be unproblematically analysed in terms of belief, and that the actual details of instantiation do not matter; appraisal theory, on the other hand, recognises the need to understand, in Scherer's terms, "the precise mechanisms involved."

The second problematic aspect of GENESE involves its construal of the relation between prototypes and emotion types. Simply put, it seems excessively narrow. In a critique of GENESE Greg Chwelos and Keith Oatley note: "this implementation contains the implicit assumption that each emotion corresponds to exactly one point in the appraisal space...[but] why in this huge vector space might not two or more distinctly different combinations of appraisal features elicit the same emotion."[37]

The proper response to this concern, which will lead to an important modification of the simple vector space model employed by GENESE, takes two forms. First, the question of whether it is justifiable to use a *single* prototype vector to define an emotion type is independent of this particular implementation. Chwelos and Oatley's suggestion that different appraisal patterns might elicit the same emotion could be accommodated by

---

[37] Chwelos and Oatley 1994, p. 249. To appreciate the force of this criticism consider the number of possible appraisal patterns in GENESE. Chwelos and Oatley calculate the number as 4.7×10^{11}, but it could vary depending upon whether appraisal patterns are seen as taking discrete or continuous values.

simply assigning the same emotion label to different vectors.[38] The main question here is instead theoretical: do different patterns of appraisal elicit the *same* emotion?

The traditional philosophical answer here, as embodied in the theory of formal objects, is 'yes.' Aristotle's analysis of anger, for example, held that it could be produced by distinctly different judgements, ranging from 'I have been the butt of excessive laughter,' to 'Your behaviour insults my intelligence.' This many-to-one relation was established on the grounds that these different judgements could be united in that qua tokened judgements they all fall under the judgement type '__ is a slight.' This judgement type, in turn, is *definitive* of the concept of anger. This defining relationship, as I've previously argued, has generally been construed as a logical one, where "logical" has itself been variously construed. Regardless of how this relation has been understood, however, philosophers have generally followed Aristotle in seeing the unifying semantic notion at the base of this picture as being that of a *tokened judgement falling under a type*. It is this relation that in turn is supposed to gather sets of different tokened emotional states—different in the minimal sense that they are elicited by different tokened judgements—under a *single* well defined emotion type.

In sharp contrast to this view is the one put forth by Scherer, a view with which I am sympathetic given the difficulties I have noted in unifying emotion types by detailing their logical relationships to defining superordinal categories. He notes: "...there may be as many different emotions as appraisal combinations and...we need to stop using a

---

[38] Wehrle and Scherer 1995, p. 603.

limited number of supposedly "basic" emotion labels for what is obviously a very highly differentiated gamut of rather different states."[39] The foundation for this view is Scherer's emphasis on emotion as *process*. Emotion, for Scherer, is

> ...a sequence of interrelated, synchronised changes in the states of all organismic subsystems (information processing/cognition, support/ANS, execution /motivation, action/SNS, monitoring/subjective feeling) in response to the evaluation of an external or internal stimulus event that is relevant to central concerns of the organism....In this sense, *the pattern of all synchronised changes in the different components over time constitutes the emotion* and even small differences in this pattern are expected to reflect real differences in the nature of the emotional state.[40]

The consequence of this view is that "the number of potential emotional states (as defined by the process of synchronised patterning of all components in the emotion episode time window) is virtually infinite."[41]

Such a claim, however, is faced with the obvious fact that we don't have an infinite number of labels for these states. More particularly, as Scherer himself notes, "we generally experience emotions in a discrete fashion...there seems to be a bundling of the different emotion states around a limited number of types."[42] This 'bundling' and the claim that our emotion states are potentially infinitely variable thus need to be reconciled.

A short answer here is that such bundling is a function of the interaction between our discrete emotional systems and relatively common types of situations—e.g., the "invariances" noted by Tooby and Cosmides— that tend to invoke typical evaluation

---

[39] Ibid., p. 604. Scherer's explicit arguments for this claim are found in Scherer 1984; 1986.
[40] Scherer 1994, p. 28.
[41] Ibid.
[42] Ibid., p. 27.

patterns by virtue of those situations being similar at some level of analysis.[43] Scherer notes: "Certain patters of expression (and possibly autonomic arousal) occur more frequently than others in response to certain types of structurally equivalent (in terms of underlying appraisal) situations and most languages provide convenient labels to refer to this "bunching."[44]

This approach to explaining the origin of relatively discrete emotion 'types'—however it is fleshed out— suggests in turn a new semantics for emotion type terms and a more precise understanding of the nature and function of prototypes. I will express the picture in terms of a modified vector space model. First, in this new model emotion type terms are not associated with *single* fixed prototype vectors—i.e., single appraisal patterns—but rather with *clusters* of vectors, i.e., with *volumes* in the state space of the appraisal system. These groupings are just those collections of similar appraisals that arise from the interaction of discrete emotional systems with emotionally salient, similarly structured situations. This is the *causal* explanation for why we find such groupings. We may ask, however, *what* justifies us in relating such groupings to single emotion type terms. In answering this question we come to a crucial difference between the theory of prototypes and the theory of formal objects.

As noted several times, philosophy traditionally relates sets of judgements to emotion type terms—and thereby gives unity to those terms— by virtue of those judgements falling under superordinate categories which are themselves related to type

---

[43] A fuller answer would also include an investigation of the processes underlying the production of verbal labels for mental states. See, e.g., Scherer 1994, pp. 30-31.

terms. I have argued, though, that this picture cannot be sustained. The basic reason for this, I now suggest, is that the 'formalism' adopted by proponents of formal objects requires that there exist a single, linguistically individuated description that is (1) logically related to an emotion type under which (2) all relevant tokened judgements can be placed as a matter of logic.[45] As I have argued, however, this is simply too strong a demand. What I want to suggest now is that vector space modelling provides a more adequately flexible formalism that still allows a principled way of establishing the correlation between sets of judgements—or appraisal processes—and emotion types while avoiding the classical problems of the philosophical approach.

The picture we have so far is of a grouping of appraisal processes, represented as single points—input vectors—in multidimensional 'appraisal space.' These groupings, in turn, can be explained *causally* as the interaction between relatively stable features in the organism and the environment. Considered formally, though, we can see that *what unifies such groupings of appraisal patterns is the geometric notion of clustering about a point*. That is, the 'descriptions' that are encapsulated in the input vectors that constitute a given cluster are not unified by their all being *logically* related to some superordinate description, but are rather unified by the fact that *when represented in vector space formalism they can be seen to converge upon a centre point.* A grouping of vectors—i.e., a range of patterns of appraisal—has, in effect, a 'centre of gravity.' This centre of gravity is a prototype. In the current implementation of GENESE this centre of gravity is

---

[44] Ibid.
[45] The 'formalism' referred to here is folk psychology.

represented as an *actual* vector; it is in fact the prototype vector associated with the related emotion label. However, as Scherer notes, a prototype vector, properly understood, is just a "fixed vector considered to be an *approximation to the central tendency*" of a particular grouping of vectors.[46] Strictly speaking, then, emotion type terms are only *indirectly* related to prototypes—i.e., prototypical appraisal patterns—and this by the fact that these prototypes serve to unify the *groupings* of appraisal patterns to which emotion types are properly related.

Emotion type terms on this model are thus quite 'loose'—Scherer uses the term "modal emotions"—in that they are not definitionally related to any *single* category or description. They are 'unified,' however, in the sense that each type is related to some prototypical appraisal pattern in the way just described. There is thus no demand here, as there is in the theory of formal objects, that there be any sort of logical, definitional relationship between an emotion type and its prototypical appraisal pattern. Dropping this demand should, in turn, help avoid the problems that it has traditionally posed.

What I want to suggest now is that appraisal theory, construed as I have suggested and expressed in the formalism sketched above, is a good candidate for the level of description that I earlier argued philosophy should work toward. On the metric of theories introduced at the beginning of this chapter this version of appraisal theory looks very much like structural psychological realism, the second option. That is, considered generally, appraisal theory constitutes a description of the relationship between emotion

---

[46] Wehrle and Scherer 1995, p. 604.

and cognition. Of course, the details of this description—the categories that it picks out as being operative in the production of our various emotions—remain to be worked out. I'll say more about this below. The *form* of this description, however, is meant to reflect the structure of the information processing capacities of the various emotional systems. That is, each hypothetical dimension in appraisal theory's state-space model is intended to represent a category that is computed by some aspect of an emotional system in the course of producing an emotion. However, unlike the first option considered by Clark—strong propositional realism—the terms of this description—i.e., the names we choose for our dimensions—are explicitly understood to be *abstractions*. When we say an emotional system has the capacity to recognise the category "novelty"—or "predator present," "conspecific expression of fear," or "reward"— we do not expect that it tokens some isomorphic symbol *novelty** in the subject's language of thought. We instead understand that "predator present" is an abstract gloss on the category actually computed—e.g., "sinusoidal movement"— by the cognitive capacity under consideration. And again, we ascend to this *particular* abstract term because it captures the significance of sinusoidal movement for the subject, and further, helps make sense of the (likely) subsequent behaviour of that subject. Any further terms that we choose in constructing our more complete description of the relations between emotion and cognition must thus perform these same functions.

Program Explanations

The basic claim of this chapter can perhaps be sharpened by comparison with an analogous set of issues concerning the relationships between the various discrete levels of description used to characterise connectionist models of cognition. I borrow here Andy Clark's description of these levels.[47]

Low-level descriptions include

1) The numeric specification of weights and activation-passing rules, and
2) Subsymbolic interpretations of the activity of processing units.

High-level descriptions include

3) The partitioning trees created by performing a cluster analysis on a network,
4) Descriptions that use the constructs of classical AI (e.g., "schema," "production," and so on), and
5) The ordinary conceptual-level descriptions of common-sense belief and desire psychology.

Supposing for the moment that human cognition is implemented in connectionist systems, the question posed by the identification of these different levels of explanation seems straightforward. Which level of description affords the correct formal account of cognition?

One potential answer, favoured by eliminativists, runs as follows. The proper account of cognition will identify just those elements in a cognitive system that are causally efficacious. In a connectionist system the causally efficacious states are the numerically characterised connection weights between nodes and the activation levels of those nodes. The behaviour of such systems at this level can thus be fully characterised mathematically by equations that describe the evolution of the system through time. It is

---

[47] Clark 1989, p. 188.

this level of description, in short, that identifies and describes the causally efficacious states of a connectionist system. Considered as an account of cognition, however, the purely mathematical nature of this level of analysis seems to place it too far from anything that could be a complete account of cognition, for such an account must seemingly make some reference to semantic content. This problem is solved by moving up one level of analysis to the subsymbolic. Roughly, this level works by placing a semantic analysis on those elements which are only mathematically characterised at the lower level. It "rather precisely interprets the numerical specification of an activation vector by associating the activation of each unit with a content."[48] The precise nature of this level of description is controversial, however, for it is sometimes claimed that the semantics of this level differ from those of the upper levels in a way that makes direct translation impossible. I will avoid this issue, as it is not really important here. The important point is rather that connectionists and eliminativists identify the subconceptual as the most basic and precise level at which an adequate description of a cognitive system may be pitched: "Complete, formal and precise descriptions of the intuitive (i.e., connectionist) processor are generally tractable not at the conceptual level, but only at the subconceptual level."[49] The further argument is then made by the eliminativists that since higher level descriptions do not pick out causally efficacious states they are explanatorily inert. They argue that those higher explanations, folk psychology in particular, should be eliminated in that they are 'technically mistaken.'

---

[48] Clark 1989, p. 189.
[49] Smolensky 1988, p.6. Quoted in Clark 1989, p. 189.

Against this general claim Clark argues there can exist legitimate causal explanations that do not involve the description of causally efficacious states. Here Clark draws on Frank Jackson and Philip Pettit's distinction between *program* and *process* explanations, a distinction best grasped through example. Jackson and Pettit observe:

> We may explain the conductor's annoyance at a concert by the fact that *someone* coughed. What will have *actually* caused the conductor's annoyance will be the coughing of some particular person, Fred, say; but when we say that it was someone's coughing that explains why the conductor was annoyed, we are thinking of someone's coughing as Fred's coughing or Mary's coughing or Harry's coughing or..., and saying that any one of these disjuncts would have caused the conductor's annoyance—it did not have to be Fred.[50]

As Clark points out, a more 'accurate' explanation of the conductor's annoyance—one that cited *Fred's* coughing—would, in an important sense, be less powerful than an explanation that cited *someone's* coughing since it would obscure the fact that "any of a whole range of members of the audience coughing would have caused annoyance in the conductor."[51]

Another example. Consider two electrons A and B that are acted upon by the forces Fa and Fb such that A subsequently accelerates at the same rate as B. Jackson and Pettit note:

> We explain the fact that electron A accelerates at the same rate as B in terms of the force acting on A being the same in magnitude as that acting on B. But this sameness in magnitude is quite invisible to A...and does not make A move off more or less briskly; what determines the rate at which A accelerates is the magnitude of Fa, not that magnitude's relationship to another force altogether.[52]

---

[50] Jackson and Pettit 1988, p. 394
[51] Clark 1989, p. 197.
[52] Jackson and Pettit 1988, p. 393.

Here, "sameness" *explains* A and B's moving off at the same rate, though it plays no role in the causal production of that result, because it identifies that property which is common to a range of cases where the causally efficacious particulars could differ while producing the same result: "Thus, in this case what actually produces the result that the accelerations are the same is both forces being of magnitude five. But both forces being of magnitude six, or seven, or...instead would equally have produced the result that the electrons' accelerations were the same."[53] By citing "sameness" in our explanation of A and B's behaviour we thereby capture this fact which would otherwise be invisible in an account that only cited the causally efficacious factors in the situation. The general point, then, is that higher level explanations that do not cite such features may nonetheless capture interesting generalisations that would not be seen at lower levels of explanation that cite only the causally efficacious features of the event being considered.

These sorts of examples thus support an important distinction between two types of explanation. A *process explanation* is any explanation that "cites the very features that are efficacious in a particular case or range of cases."[54] In contrast, a *program explanation* is any explanation that "highlights a common feature of a range of cases...but abstracts away from the causally active features of a particular case...."[55] In the above examples the properties that are highlighted—the *sameness* of the forces acting on the electrons, and *someone's* coughing—are said to "causally program" the results that they explain. They do this, in turn, by *ensuring*—as opposed to causing— that *some* set

---

[53] Ibid.
[54] Clark 1989, p. 189.

of causally efficacious factors obtains, though they do not specify which set did in fact obtain. That is, any set of causally efficacious factors that fall under the description tendered in an accurate program explanation will necessarily be capable of causally producing the results that the property was invoked to explain. Of course, some program explanations will be more informative than others.[56] Consider, for example, an account of the conductor's annoyance that explained it by saying that 'someone in a tuxedo coughed.' While it is true that any particular falling under this description—i.e., anyone wearing a tuxedo who coughed—would have caused the conductor to become annoyed, this particular program explanation is not maximally informative since the property it invokes to explain the conductor's annoyance fails to collect together other particulars—say those who wore suits and coughed—that would have had the same effect.

Turning now to explanations of human behaviour we can see that, strictly considered, a *pure* process explanation of some piece of behaviour will only pick out the neurophysiological aspects of the states and processes that led to that behaviour. Thus if cognition is implemented in the human brain by connectionist systems, a pure process explanation will only specify the numerically characterised connection weights between nodes, and the activation levels of those nodes, since as noted above these are the only features of a connectionist system that are causally efficacious. Just as in the above examples, however, such a narrow description will clearly place sharp limits on our

---

[55] Ibid., p. 198.
[56] For a fuller discussion see Jackson and Pettit 1988, p. 396.

capacity to make interesting and important generalisations, since as we have seen such generalisations are sometimes only available at higher levels of description.

Clark points out, for example, the explanatory virtues of cluster analysis when applied to a connectionist system like NETtalk, which models the process of converting written words into speech. Cluster analysis, in short, is an attempt to discern the divisions in the state space of a connectionist system that the system has made in the course of learning a new task. Clark notes:

> It is an important fact about cluster analysis that networks that come to embody different connection weights may have identical cluster analyses. [For example]...versions of NETtalk that begin with different random distributions of weightings on the hidden units will, after training, make the same partitions but by means of different arrangements of weights on the individual connections. Now consider a particular cognitive domain, say converting text to phonemes. Isn't it a legitimate psychological fact that only certain systems can successfully negotiate that domain? And don't we want some level of properly psychological, or cognitive, explanation with the means to group such systems together and to make some generalisations about them (e.g., that such systems will be prone to certain illusions)? Cluster analysis is the very tool we need.[57]

Cluster analysis works here as an informative program explanation because the terms and constructs that it employs allow the expression of important commonalties and patterns in the systems being examined. For example, its central theoretical construct—clusters in state space—allows us to express the fact that different activation patterns can produce the same result. It does this by providing a form of representation that shows *explicitly* how those different patterns must be related—they must fall within the same cluster.[58]

---

[57] Clark 1989, p. 199.

[58] Imagine how difficult it would be to express such patterns if all relevant activation vectors were simply listed in numerical form. The lesson here is that different forms of representation allow certain facts to be

Again, however, clusters *per se* are not causally efficacious features of a connectionist system. Clearly, however, this does not stop them from figuring in informative explanations.

A similar defence can be made for ever higher levels of description, folk psychology included.[59] Always, though, any description tendered must ultimately justify itself by "offering a terminology that groups various systems into psychologically interesting equivalence classes that are unmotivatable if we restrict ourselves to [lower levels of description]."[60] Regarding folk psychology in particular we should thus expect that the constructs it employs similarly work to pick out interesting equivalence classes: "The belief construct must earn its keep by grouping together creatures whose gross behaviours really do have something important in common (e.g., all those likely to harm me because they *believe* I am a predator)."[61] Of course, the belief construct *does* often earn its keep in this way. As argued in the previous chapters, however, belief sometimes

---

expressed more easily than others.

[59] Dennett offers a nice example here that brings out especially clearly the virtue and function of program explanations. He asks us to imagine the arrival on earth of Martians of surpassing intelligence. They are perfect Laplacean physicists, able to conceive of humans as swarms of particles and make flawless predictions about our behaviour from this physical stance alone. Dennett then asks us to compare the explanatory virtues and predictive powers of this stance with one that invokes intentional descriptions—i.e., the "Intentional Stance." He notes: "Our imagined Martians might be able to predict the future of the human race by Laplacean methods, but if they did not also see us as intentional systems, they would be missing something perfectly objective: the *patterns* in human behaviour that are describable from the intentional stance, and only from that stance, and that support generalisations and predictions. Take a particular instance in which the Martians observe a stockbroker deciding to place an order for 500 shares of General Motors. They predict the exact motions of his fingers as he dials the phone and the exact vibrations of his vocal cords as he intones his order. *But if the Martians do not see that indefinitely many different patterns of finger motions and vocal cord vibrations...could have substituted for the actual particulars without perturbing the subsequent operation of the market, then they have failed to see a real pattern in the world they are observing*" (Dennett 1987, pp. 25-26).

[60] Clark 1989, p. 198.

[61] Ibid., p. 201.

fails in this regard. Just like the explanation of the conductor's annoyance that cites 'someone in a tuxedo's coughing,' the demands of the belief construct are often such that it fails to collect together importantly different mental states that nonetheless are capable of producing the same resultant emotion. The various experiments cited in the second chapter were intended to prove just this point.

Putting matters in this way now allows me to restate more generally the conclusion of my arguments in the last two chapters, and the claims that I have made for appraisal theory and vector space formalism in this one. In short, I have argued that folk psychology, as loosely 'formalised' in the philosophical theory of formal objects, often fails to establish interesting, principled equivalencies between emotional and cognitive states. More particularly, I have argued that classes of beliefs or judgements, defined by their logical relation to formal objects, do not stand in any particularly interesting relationship to well-bounded emotion types. Conversely, my comments on appraisal theory and the vector space formalism employed by Scherer have been an attempt to sketch a form and level of description that is more likely to be capable of expressing and explaining interesting cognition-emotion equivalencies. This level of description has as its central cognitive construct the functional notion of 'appraisal,' and the semantics of this construct are provided by the multi-dimensional state space model discussed here. The identity of the various appraisal dimensions that constitute this model, in turn, are intended to be informative glosses on the categories actually operant in the production of an emotion, glosses that as previously explained allow us to see the wider significance of those categories for the subject.

Conclusion

Considered generally, the basic goal of this chapter has been to sketch a program that allows us to begin to bring together the unique insights of the three approaches to emotion that I have considered in this thesis: formal object theory, appraisal theory, and the empirical investigation of emotion systems. Hopefully, some of the links between these approaches have been made clear. Both formal object and appraisal theory, for example, can be understood as seeking a formal *descriptive* account of the relationship between emotion and cognition. They work toward this end by detailing, for a particular emotion type, the general description under which a subject's causally efficacious mental states must *in effect* place a stimulus if that emotion is to be produced. I have argued, however, that appraisal theory, in the form sketched above, is likely in a better position to express this formal account. Part of the reason for this is that its central construct—appraisal—is more liberal with regard to the causally efficacious states that can be collected under it. This liberality is especially valuable now in light of the empirical evidence emerging from the investigation of emotional systems that shows that many of our emotions can be produced by a range of states that are importantly unlike beliefs. And while more liberal in this regard, appraisal theory has the further advantage that it still explicitly recognises the need to 'descend' into the *causal* cognitive sciences so as to better understand the actual nature of the states and processes that are confounded under the appraisal construct. We must descend to this level of explanation, in part, because it is the *nature* of these elements to act as one of the central constraints on the construction of the specific appraisal dimensions that constitute appraisal theory's formal model.

Appendix A[1]

GENESE

Patterns of Stimulus Evaluation Checks Predicted to Differentiate 14 Major Emotions

	Enjoyment/ Happiness	**Elation/ Joy**	**Displeasure/ Disgust**	**Contempt/ Scorn**	**Sadness/ Dejection**	**Despair**	**Anxiety/ Worry**
Novelty							
Suddenness	Low	Hi/med	Open	Open	Low	High	Low
Familiarity	Open	Open	Low	Open	Low	Very low	Open
Predictability	Medium	Low	Low	Open	Open	Low	Open
Intrinsic Pleasantness	High	Open	Very low	Open	Open	Open	Open
Goal Significance							
Concern Relevance	Open	Self/relationship	Body	Relationship/order	Open	Open	Body/self
Outcome Probability	Very High	Very high	Very high	High	Very high	Very high	Medium
Expectation	Consonant	Open	Open	Open	Open	Dissonant	Open
Conduciveness	Conducive	Very conducive	Open	Open	Obstruct	Obstruct	Obstruct
Urgency	Very Low	Low	Medium	Low	Low	High	Medium
Coping Potential							
Cause: Agent	Open	Open	Open	Other	Open	Other/nature	Other/nature
Cause: Motive	Intent	Chance/intention	Open	Intention	Chance/negligence	Chance/negligence	Open
Control	Open	Open	Open	High	Very low	Very low	Open
Power	Open	Open	Open	Low	Very low	Very low	Low
Adjustment	High	Medium	Open	High	Medium	Very low	Medium
Compatibility Standards							
External	Open	Open	Open	Very low	Open	Open	Open
Internal	Open	Open	Open	Very low	Open	Open	Open

[1] Adapted from Scherer 1993, pp. 338-339.

	Fear	**Irritation/ Cold Anger**	**Rage/ Hot Anger**	**Boredom/ Indifference**	**Shame**	**Guilt**	**Pride**
Novelty							
Suddenness	High	Low	High	Very low	Low	Open	Open
Familiarity	Open	Open	Low	High	Open	Open	Open
Predictability	Low	Medium	Low	Very high	Open	Open	Open
Intrinsic Pleasantness	Low	Open	Open	Open	Open	Open	Open
Goal Significance							
Concern Relevance	Body	Order	Order	Body	Self	Relationships/order	Self
Outcome Probability	High	Very high	Very high	Very high	Very high	Very high	Very high
Expectation	Dissonant	Open	Dissonant	Consonant	Open	Open	Open
Conduciveness	Obstruct	Obstruct	Obstruct	Open	Open	High	High
Urgency	Very High	Medium	Medium	Low	High	Medium	Low
Coping Potential							
Cause: Agent	Other/Nature	Open	Other	Open	Self	Self	Self
Cause: Motive	Open	Intent/negligence	Intent	Open	Intent/negligence	Intention	Intention
Control	Open	High	High	Medium	Open	Open	Open
Power	Very low	Medium	High	Medium	Open	Open	Open
Adjustment	Low	High	High	High	Medium	Medium	High
Compatibility Standards							
External	Open	Low	Low	Open	Open	Very low	High
Internal	Open	Low	Low	Open	Very low	Very low	Very high

Bibliography

Adolphs, Ralph, Daniel Tramel, Hannah Damasio, and Antonio. Damasio. 1994. "Impaired Recognition of Emotion in Facial Expressions Following Bilateral Damage to the Human Amygdala." *Nature*, vol. 372, no. 15: 669-672.

Alston, William. 1967. "Emotion and Feeling," in Paul Edwards, ed., *Encyclopedia of Philosophy*. New York: MacMillan.

Aristotle. *Nichomachean Ethics*, trans. J. A. K. Thomson. London: Penguin Books (1976).

_____. *Rhetoric*, trans. Theodore Buckley. Amherst, NY: Prometheus (1995).

Armony, Jorge L., and Joseph LeDoux. 2000. "How Danger Is Encoded: Toward a Systems, Cellular, and Computational Understanding of Cognitive-Emotional Interactions in Fear," in Michael Gazzaniga, ed., *The New Cognitive Neurosciences, 2nd edition.* Cambridge: MIT Press.

Averill, James. 1994. "It's a Small World, But a Large Stage," in Paul Ekman and Richard Davidson, eds., *The Nature of Emotion: Fundamental Questions.* Oxford: Oxford University Press.

Bartky, Sandra. 1990. *Femininity and Domination: Studies in the Phenomenology of Opression.* New York: Routledge.

Bechara, Antoine, Daniel Tramel, Hanna Damasio, Ralph Adophs, Charles Rockland, and Antonio Damasio. 1995. "Double Dissociation of Conditioning and Declarative Knowledge Relative to the Amygdala and Hippocampus in Humans." *Science* 269: 1115-1118.

Bedford, Errol. 1957. "Emotion." *Proceeding of the Aristotelian Society* 57:281-304. Reprinted in *Essays in Philosophical Psychology*, ed. D. Gustafson. Garden City, NY: Doubleday (1964).

Brothers, Leslie, Brian Ring, and Arthur Kling. 1990. "Response of Neurons in the Macaque Amygdala to Compex Social Stimuli." *Behavioural Brain Research* 41: 199-213.

Browne, Janet. 1985. "Darwin and the Expression of the Emotions," in David Kohn, ed., *The Darwinian Heritage.* Princeton: Princeton University Press.

Buck, Ross. 1986. "The Psychology of Emotion," in Joseph LeDoux and William Hirst, eds., *Mind and Brain: Dialogues on Cognitive Neuroscience.* New York: Cambridge University Press..

Burkhardt, Richard. 1985. "Darwin on Animal Behaviour and Evolution," in David Kohn, ed., *The Darwinian Heritage.* Princeton: Princeton University Press.

Campbell, Sue. 1997. *Interpreting the Personal: Expression and the Formation of Feelings.* Ithaca: Cornell University Press.

Caston, Victor. 1998. "Aristotle and the Problem of Intentionality." *Philosophy and Phenomenological Research*, vol. LVIII, no. 2: 249-297.

Churchland, Paul. 1979. *Scientific Realism and the Plasticity of Mind.* Cambridge: Cambridge University Press.

Chwelos, Greg, and Keith Oatley. 1994. "Appraisal, Computational Models, and Scherer's Expert System." *Cognition and Emotion* 8 (3): 245-257.

Clark, Andy. 1989. *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* Cambridge: MIT Press.

_____. 1997. *Being There: Putting Brain, Body, and World Together Again.* Camridge: MIT Press.

_____. 1998. "Time and Mind." *The Journal of Philosophy*, vol. XCV, no. 7: 354-376.

Cosmides, Leda, and John Tooby. 2000. "Evolutionary Psychology and the Emotions," in T. Dalgleish and M. Power, eds., *Handbook of Emotions, 2nd ed.* NY: Guilford Press.

Damasio, Antonio. 1994. *Descartes Error: Emotion, Reason, and the Human Brain.* New York: Grosset/Putnam.

Darwin, Charles. 1872. *The Expression of the Emotions in Man and Animals, 3rd ed., Introduction, Afterword and Commentaries by Paul Ekman.* Oxford: Oxford University Press (1998).

Davidson, Donald. 1980. *Essays on Actions and Events.* Oxford: Clarendon Press.

Davis, Michael. 1992. "The Role of the Amygdala in Fear and Anxiety. *Annual Review of Neuroscience* 15: 353-375.

Dennett, Daniel. 1969. *Content and Consciousness.* London: Routledge & Kegan Paul.

_____. 1977. "A Cure for the Common Code?" *Mind* 86: 265-280.

_____. 1987. *The Intentional Stance.* Cambridge: Bradford.

_____. 1995. *Darwin's Dangerous Idea.* New York: Simon & Schuster.

Descartes, René. 1985. *The Philosophical Writings of Descartes, Vol. 1*, trans. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press.

De Sousa, Ronald. 1990. *The Rationality of Emotion.* Cambridge: MIT Press.

_____. 1996. "Prefrontal Kantians. A Review of *Descartes Error: Emotion, Reason and the Human Brain* by Antonio R. Damasio." *Cognition and Emotion* 10 (3): 329-333.

Dewey, John. 1894. "A Theory of Emotion." *Psychological Review* 1: 553-69; 2: 13-52. Reprinted in *John Dewey: The Early Work,* 1882-1898, vol. 2. Carbondale: Southern Illinois University Press, 1967.

Dicks, Dennis, Ronald E. Myers, and Arthur Kling. 1969. "Uncus and Amygdala Lesions: Effects on Social Behaviour in the Free Ranging Rhesus Monkey." *Science* 165: 69-71.

Dolan, Raymond J. 2000. "Emotional Processing in the Human Brain Revealed through Functional Neuroimaging," in Michael Gazzaniga, ed., *The New Cognitive Neurosciences, 2nd edition.* Cambridge: MIT Press.

Duclos, Sandra, *et al.* 1989. "Emotion-specific Effects of Facial Expressions and Postures on Emotional Experience." *Journal of Personality and Social Psychology,* vol. 57, no. 1: 100-108.

Ekman, Paul. 1980. "Biological and Cultural Contributions to Body and Facial Movement in the Expression of Emotions," in Amélie Rorty, ed., *Explaining Emotions.* Berkeley: University of California Press.

_____. 1992. "An Argument for Basic Emotions." *Cognition and Emotion* 6: 169-200.

_____. 1994. "Antecedent Events and Emotions Metaphors," in Paul Ekman and Richard Davidson, eds., *The Nature of Emotion: Fundamental Questions.* Oxford: Oxford University Press.

Ekman, Paul, Richard Sorenson and Wallace Friesen. 1969. "Pan-cultural Elements in Facial Displays of Emotion." *Science* 164: 86-88.

Ekman, Paul, R. Levenson, and Wallace Friesen. 1983. "Autonomic Nervous System Activity Distinguishes Among Emotions." *Science* 221: 1208-1210.

Ekman, Paul, and Richard Davidson, eds. 1994. *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

Ellsworth, Phoebe. 1994. "Some Reasons to Expect Universal Antecedents of Emotion," in Paul Ekman and Richard Davidson, eds., *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

Fodor, Jerry. 1975. *The Language of Thought*. New York: Cromwell and Company.

_____. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge: MIT/Bradford.

Frijda, Nico. 1993. "Appraisal and Beyond." *Cognition and Emotion* 7 (3/4): 225-231.

_____. 1994. "Universals in Antecedents of Emotion," in Paul Ekman and Richard Davidson, eds., *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

Gazzaniga, Michael, ed. 2000. *The New Cognitive Neurosciences, 2nd edition.* :MIT Press

Gordon, Robert. 1987. *The Structure of Emotions*. New York: Cambridge University Press.

Griffiths, Paul. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.

Halligan, Peter. 1998. "Inability to Recognise Disgust in Huntington's Disease. *Lancet*, Feb 14.

Heining, M., A.W. Young, S.R.C Williams, C. Andrew, M.J. Brammer, J.A Gray, M.L. Phillips. 2000. "Neural Responses to Auditory and Visual Presentations of Anger, Disgust, Fear and Sadness." Poster No.: 243, 6th Annual Meeting of the ORGANIZATION FOR HUMAN BRAIN MAPPING Conference on Functional Mapping of the Human Brain. In *NeuroImage*, vol. II, no. 5, part 2.

Hume, David. 1739. *A Treatise of Human Nature*, ed. L. A. Selby-Bigge. Oxford: Clarendon Press (1964).

Iidaka, Tetsuya, Tetsuhito Murata, Masao Omori, Hirotaka Kosaka, Tomohisa Okada, Norihiro Sadato, Yoshiharu Yonekura. 2000. "Functional MRI Study of Neural Network Involved in Emotional Face Processing." Poster No.:230, 6th Annual Meeting of the ORGANIZATION FOR HUMAN BRAIN MAPPING Conference on Functional Mapping of the Human Brain. In *NeuroImage,* vol. II, no. 5, part 2.

Izard, Carroll E. 1993. "Four Systems for Emotion Activation: Cognitive and Noncognitive Processes." *Psychological Review*, vol. 100, no. 1: 68-90.

Jackson, Frank, and Philip Pettit. 1988. "Functionalism and Broad Content." *Mind* 97, no. 387: 381-400.

James, William. 1893. *The Principles of Psychology.* New York: Dover (1950).

Kenny, Anthony. 1963. *Action, Emotion and Will.* London: Routledge & Kegan Paul.

Lavadas, E., D. Cimatti, M. Del Pesce and G. Tuozzi. 1993. "Emotional Evaluation With and Without Conscious Stimulus Identification: Evidence from a Split Brain Patient." *Cognition and Emotion* 7 (1): 95-114.

Lazarus, Richard. 1982. "Thoughts on the Relations Between Emotion and Cognition." *American Psychologist*, vol 37, no 2: 1019-1024.

_____. 1994. "Universal Antecedents of the Emotions," in Paul Ekman and Richard Davidson, eds., *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

LeDoux, Joseph. 1986. "The Neurobiology of Emotion," in Joseph LeDoux and William Hirst, *Mind and Brain: Dialogues on Cognitive Neuroscience.* New York: Cambridge University Press.

_____. 1993. "Cognition versus Emotion, Again - This Time in the Brain: A Response to Parrott and Schulkin." *Cognition and Emotion* 7 (1): 61-64.

_____. 1996. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life.* New York: Simon & Schuster.

Leonard, C.M., E.T. Rolls, F.A.W. Wilson, and G.C. Baylis. 1985. "Neurons in the Amygdala of the Monkey with Responses Selective for Faces." *Behavioural Brain Research* 15: 159-176.

Levenson, R., Paul Ekman, and Wallace Friesen. 1990. "Voluntary Facial Action Generates Emotion-Specific Autonomic Nervous System Activity." *Psychophysiology* 27: 363-384.

Leventhal, Howard, and Klaus Scherer. 1987. "The Relationship of Emotion to Cognition: A Functional Approach to a Semantic Controversy." *Cognition and Emotion* 1 (1):3-28.

Lyons, William. 1980. *Emotion*. Cambridge: Cambridge University Press.

Millikan, Ruth. 1984. *Language, Thought and Other Biological Categories: New Foundations for Realism*. Cambridge: MIT/Bradford.

Morris, J.S., C.J. Frith, D.I. Perret, *et al*. 1996. A Differential Neural Response in the Human Amygdala to Fearful and Happy Facial Expressions." *Nature* 383:812-15.

Murphy, Sheilah T., and R.B. Zajonc. 1993. "Affect, Cognition, and Awareness: Affective Priming With Optimal and Suboptimal Stimulus Exposures." *Journal of Personality and Social Psychology*, vol. 64, no. 5:723-739.

Murphy, Sheilah T., R.B. Zajonc, and Jennifer Monahan. 1995. "Additivity of Nonconscious Affect: Combined Effects of Priming and Exposure." *Journal of Personality and Social Psychology*, vol. 69, no. 4: 589-602.

Nesse, Randolph, and George Williams. 1994. *Why We Get Sick: The New Science of Darwinian Medicine*. New York: Random House.

Nussbaum, Martha. 1994. *The Therapy of Desire: Theory and Practice in Hellenistic Ethics*. Princeton: Princeton University Press.

Ono, Taketoshi, and Hisao Nishijo. 2000. "Neurophysiological Basis of Emotion in Primates: Neuronal Responses in the Monkey Amygdala and Anterior Cingulate Cortex," in Michael Gazzaniga, ed., *The New Cognitive Neurosciences, 2nd edition*. Cambridge: MIT Press.

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

Parrott, Gerrod W., and Jay Schulkin. 1993a. "Neurophysiology and the Cognitive Nature of the Emotions." *Cognition and Emotion* 7 (1): 43-59.

Parrott, Gerrod W., and Jay Schulkin. 1993b. "What Sort of System Could An Affective System Be? A Reply to LeDoux." *Cognition and Emotion* 7 (1): 65-69.

Pfeifer, Rolf. 1988. "Artificial Intelligence Models of Emotion," in *Proceedings of the NATO Advanced Research Workshop on Cognitive Science Perspectives on Emotion, Motivation and Cognition*. Dordrecht: Kluwer Academic.

Philips, Mary, Lea Williams, Andrew Young, Christopher Andrew, Ed Bullmore, Michael Brammer, Steve Williams, Michael Morgan, and Jeffrey Gray. 2000. Differential Neural Responses to Overt and Covert Presentations of Facial Expressions of Fear and Disgust. Poster No.:232, 6th Annual Meeting of the ORGANIZATION FOR HUMAN BRAIN MAPPING Conference on Functional Mapping of the Human Brain. In *NeuroImage*, vol. II, no. 5, part 2.

Picard, Rosalind. 1997. *Affective Computing*. Cambridge: MIT Press.

Plato. *Collected Dialogues*, Edith Hamiliton and Huntington Cairns, eds. Princeton, NJ: Princeton University Press (1961).

Quine, W.V. 1951. "Two Dogmas of Empiricism." *The Psychological Review* 60. Reprinted in Quine, W.G., *From a Logical Point of View*. New York: Random House (1963).

Rey, Georges. "Functionalism and the Emotions," in Amélie Rorty, ed., *Explaining Emotions*. Berkeley: University of California Press.

Rorty, Amélie, ed. 1980. *Explaining Emotions*. Berkeley: University of California Press.

Roseman, Ira, Ann Aliki Antoniou and Paul Jose. 1996. "Appraisal Determinants of Emotions: Constructing a More Accurate and Comprehensive Theory." *Cognition and Emotion* 10 (3): 241-277.

Schachter, Stanley, and Jerome E. Singer. "Cognitive, Social, And Psycholgical Determinants of Emotional State." *Psychological Review* 69 (5): 379-399.

Scherer, Klaus. 1984. "On the Nature and Function of Emotion: A Component Process Approach," in Klaus Scherer and Paul Ekman, eds., *Approaches to Emotion*. Hillsdale: Erlbaum Press.

_____. 1986. "Vocal Affect Exression: A Review and a Model for Future Research." *Psychological Bulletin*, vol. 99, no. 2: 143-165.

_____. 1988. "Criteria for Emotion-Antecedent Appraisal: A Review," in Vernon Hamilton, Gordon Bower, and Nico Frijda, eds., *Cognitive Perspectives on Emotion and Motivation (NATO Asi Series, Series D: Behavioural and Social Sciences, vol. 44)*. Boston: Kluwer.

_____. 1993. "Studying the Emotion-Antecedent Appraisal Process: An Expert System Approach." *Cognition and Emotion* 7 (3/4): 325-355.

_____. 1993a. "Neuroscience Projections to Current Debates in Emotion Psychology." *Cognition and Emotion* 7 (1): 1-41.

_____. 1994. "Toward a Concept of "Modal Emotions," in Paul Ekman and Richard Davidson, eds., *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

_____. 1994a. "Evidence for Both Universality and Cultural Specificity of Emotion Elicitation," in Paul Ekman and Richard Davidson, eds., *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

_____. 1999. "Appraisal Theory," in T. Dalgleish and M. Power, eds., *Handbook of Cognition and Emotion*. New York: John Wiley & Sons.

Seamon, J. G., N. Brody, and D.M. Kauf. 1983. "Affective Discrimination of Stimuli That Are Not Recognized: Effects of Shadowing, Masking, and Cerebral Laterality." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9: 544-555.

Solomon, Robert. 1980. "Emotions and Choice," in Amélie Rorty, ed., *Explaining Emotions*. Berkeley: University of California Press.

_____. 1993. *The Passions: Emotions and the Meaning of Life*. Indianapolis: Hackett.

Sterelny, Kim. 1990. *The Representational Theory of Mind*. Oxford: Basil Blackwell.

Stich, Stephen. 1978. "Beliefs and Subdoxastic States." *Philosophy of Science* 45: 499-518.

_____. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge: Bradford.

Teasdale, John. 1999. "Multi-level Theories of Cognition-Emotion Relations," in T. Dalgleish and M. Power, eds., *Handbook of Cognition and Emotion*. New York: John Wiley & Sons.

Thalber, Irving. 1977. *Perception, Emotion and Action*. Oxford: Basil Blackwell.

Tooby, John, and Leda Cosmides. 1990. "The Past Explains the Present: Emotional Adaptions and the Structure of Ancestral Environments." *Ethology and Sciobiology* 11: 375-424.

Tversky, Barbara, and Kathleen Hemenway. 1983. "Categories of Environmental Scenes." *Cognitive Psychology* 15:121-149.

Wehrle, Thomas, and Klaus Scherer. 1 995. "Potential Pitfalls in Computational Modelling of Appraisal Processes: A Reply to Chwelos and Oatley." *Cognition and Emotion* 9 (6): 599-616.

Winton, W. 1986. "The Role of Facial Response in Self-reports of Emotion: A Critique of Laird." *Journal of Personality and Social Psychology*, vol. 50: 808-812.

Zajonc, Robert. 1980. "Feeling and Thinking: Preferences Need No Inferences." *American Psychologist*, vol. 35,, no. 2: 151-175.

_____. 1984. "On Primacy of Affect." in Klaus Scherer and Paul Ekman, eds., *Approaches to Emotion*. Hillsdale: Erlbaum Press.