Emotions

An Essay in Aid of Moral Psychology

ROBERT C. ROBERTS
Baylor University

CAMBRIDGE
UNIVERSITY PRESS

Contents

1

Studying Emotions

1.1. WHAT THIS BOOK IS ABOUT

Anthony Trollope comments about an unsavory character who looms large in his novel *The Prime Minister* (Chapter 58):

> The abuse which was now publicly heaped on the name of Ferdinand Lopez hit the man very hard; but not so hard perhaps as his rejection by Lady Eustace. That was an episode in his life of which even he felt ashamed, and of which he was unable to shake the disgrace from his memory. He had no inner appreciation whatsoever of what was really good or what was really bad in a man's conduct. . . . In a sense he was what is called a gentleman. He knew how to speak, and how to look, how to use a knife and fork, how to dress himself, and how to walk. But he had not the faintest notion of the feelings of a gentleman. He had, however, a very keen conception of the evil of being generally ill spoken of.

Without directly mentioning any of Lopez's actions, Trollope here unmistakably sketches a man of momentous moral defects, just by indicating his patterns of emotional responsiveness – that he is more ashamed of being rejected by a classy female adventurer than of being the object of public moral opprobrium, but not at all ashamed of his shameful deeds. His lack of appreciation for good and bad action, suggests Trollope, is due to his emotional unresponsiveness to actions in moral terms (notice how Trollope mixes descriptions of Lopez's emotional dispositions with cognitive ascriptions like "no inner appreciation," "not the faintest notion," "a very keen conception"). The structure of his emotions explains why he does so much evil, why he has so little moral understanding, and why his life and the lives of those he touches closely are so miserable.

The involvement of emotions in what may be broadly termed the "moral" character of our lives is pervasive and deep. Because emotions are often impulses to act, their quality strongly affects the quality of what we do. Those who are prone to strong and inappropriate fear and anger tend to act and behave in a certain set of familiar ways, while compassion and the emotions

of friendship incline people to actions of another kind. These two sorts of emotional tendencies, and many others, may coexist in a single person, thus making people complex and morally puzzling. But emotions are not *just* "causes" of actions; they may also determine the identity of our actions. The very "same" action of shoving a person into a ditch may be done from anger at the shoved person or fear for her life; in the first case the agent is *getting revenge* (let us say), and in the second he is *protecting against danger*. What I have said has already suggested that our character or personality is in large part a disposition to be affected in one set of ways or another: One who is regularly angered by trivial offenses to his private person but seldom or never by significant offenses against others or against the public good is a mean-spirited person. Someone who rejoices in the flourishing of family and friends, for their sake, has a nobler character than one who is unaffected by their weal, or who is affected by it, not for their sake but, say, for the sake of his own convenience. To be emotionally unsusceptible to another's well- or ill-being for the other's sake is to be incapable of friendship with that other, on at least one conception of friendship; the most important relationships of our lives are constituted, in large part, by our dispositions to react with specific emotions to the other and his vicissitudes. Besides these connections to action, character, and relationships, emotions are a kind of eye for value and the import of situations, a mode of spiritual perception that may be deep and wise, or shallow and foolish. Because of these and other types of importance, certain regular patterns of emotional response are characteristic of the flourishing, mature, and "happy" human life, while alternative patterns constitute ill-function and immaturity and tend to misery.

This volume and its projected companion aim to contribute to our understanding of moral personality conceived in a broad sense of "moral," with a particular focus on the place of emotions and emotional formation in that personality. The conception of *moral* to which I refer includes not only our responses to duties and permissions, but also our happiness (which certainly does not imply always feeling good) – what kind of life, and in particular what formation of personality, and thus of relationships with others, constitutes human well-being all around. The work is divided into two parts. The present volume is on the nature of emotions and feelings and, in Chapter 4, begins to treat their connections to the moral life. The projected second volume begins with a general account of the relation of emotions to morality in my broad sense of the word, and then it offers accounts of a number of particular traits of the flourishing personality with special reference to the emotions and emotionlike states that exemplify or interact with them.

The project of understanding the good life in terms of the virtues, and the virtues in terms of their relationships to the emotions, is nothing new. Aristotle says that moral virtue is concerned with passions and actions, and in his accounts of particular virtues the passions often figure even more

prominently than the actions. The association of the virtues with the passions (many of which we would call emotions in modern English) recurs almost wherever the virtues are carefully reflected on in the history of philosophy. Thus Thomas Aquinas devotes Questions 22–48 of the first part of the second part of his *Summa Theologiæ* to a study of the passions, preparatory to his general discussion of virtue in Questions 55–67 and his detailed discussions of the virtues and vices in the second part of the second part, Questions 1–170, many of which themselves involve discussions of passions such as hope, fear, despair, joy, love, hatred, and envy. Book II of David Hume's *A Treatise of Human Nature* – "Of the Passions" – prepares the reader for Book III, in which he presents his ethics of virtue. Adam Smith's ethics, as presented in *The Theory of Moral Sentiments*, is likewise an ethics of virtue that focuses strongly on the passions. In our own period, when John Rawls turns to address justice as a trait of persons (rather than a structural feature of institutions), he finds it necessary to speak not just of dispositions to act, but of moral sentiments such as anger and guilt.[1]

This book is not a historical work, but I intend it as a contribution to this long discussion. I hope that it is in some ways a refinement of its predecessors. At any rate, it is dependent on them for direction and inspiration, as well as for the proposals that have fueled my thought, even when I disagree with them. As befits its historical location at the beginning of the 21st century, this book is more sensitive than its forebears to the possibility that neither emotions nor virtues are the same in every cultural setting, but instead vary to some extent with systems of custom, interest, and belief. While attempting to credit the diversity or potential diversity of human emotions and virtues, my discussions are also more resolutely particularistic. It seems to me that the way to study virtue is to study the virtues, and to do so rather in depth. "Virtue theory," especially in our time but also earlier, has often been long on generalizing accounts and short on careful exploration of particular virtues. Particular virtues are treated as illustrations of general theory, rather than as a fund of insight out of which any generalizations that are possible may emerge.

Accordingly, much of the second volume will study particular virtues, with special emphasis on their dynamic and internal connections with emotions and emotion dispositions. I comment on more general theoretical questions because they seem naturally to arise out of the particular discussions of virtues. Thus the method is "empirical" in the broad sense that it follows Wittgenstein's dictum, "Don't think. Look!", though lots of the looking in this sort of case is a kind of thinking – thinking about examples, as Wittgenstein's also is. Similarly, in the present volume, Chapter 3 is devoted to an extensive detailing of particular emotion types as well as of

[1] *A Theory of Justice* (Oxford: Oxford University Press, 1972), Sections 66–67 and 73–74.

emotionlike phenomena that are sometimes treated as emotions, such as surprise, amusement (at what is comical), and vanity.

1.2. THE SUPPOSED POVERTY OF CONCEPTUAL ANALYSIS

How shall we conduct an inquiry into the emotions that will serve well the study of the virtues? The methods of many disciplines have been used to study the emotions. Philosophers, from Aristotle[2] to the present, have used an approach that today would be called conceptual or philosophical analysis, one that I want to examine closely in this opening chapter because recently it has been under attack and I will argue that it is still the central approach for our purposes. But in the 19th and 20th centuries a number of other approaches have been developed. Emotions have been examined by the methods of evolutionary biology, experimental psychology, brain science, psychoanalysis and other clinical approaches, cultural anthropology, and cultural history and the history of ideas. In each case, one or another of a variety of theories forms a more or less definite background of the examination and shapes its results. For example, evolutionary biologists tend to think of emotions as behavioral response mechanisms that (at least in our evolutionary past, and in some cases also now) promote physical survival, while many anthropologists think of emotions as culturally determined patterns of experience and behavior that serve various social functions (though some anthropologists are psychoanalytic, and so stress less the determinations of culture). Brain scientists tell a rather different story about emotions, one in terms of brain circuitry and neurotransmitters, but typically lean on the evolutionary conception of emotion, while historians of the emotions may exploit psychoanalytic theory or perhaps a more cognitive-behavioral framework. In addition to these disciplines, fiction writing should be mentioned, though it is not theoretical or academic in the way the other disciplines are. Nevertheless, writers such as Jane Austen, Fyodor Dostoevsky, Leo Tolstoy, Charles Dickens, and George Eliot are very astute observers of emotions in the context of the narrative flow of human life, and are especially important for our purposes since they so often depict the emotions as expressing traits of persons' character. Most of the other disciplines focus much less on emotions that differentiate persons of one moral formation from persons of another, and seldom are emotions set in as rich a narrative context as they are in literature. A possible exception is psychoanalysis.

Conceptual analysis is an approach to the investigation of emotions that takes major clues about them from the ways people talk about the emotions in the contexts of their life. As I understand the practice and as the word "clue" suggests, it is not a purely lexicographical or syntactical/semantic

[2] See Aristotle's *Rhetoric*, especially Book II, Chapters 1–11 (1377b15–1388b30).

approach. It is not as though we could expect to find out what emotions are by looking up "emotion" and/or words like "anger" and "nostalgia" in the dictionary, nor could we expect to "analyze" such concepts merely by summarizing the various conditions under which the most proficient speakers of English ascribe emotion or anger to themselves and others. Such information about how the best speakers of English use the emotion words is an important part of conceptual analysis, but the analyst is very much in the business of *interpreting* these facts of usage. For one thing, even the best English speakers use vocabulary loosely and shiftingly, so conceptual analysis will involve normative decisions about what is the right and central usage. But beyond this, the conceptual analyst typically offers some general schema by which he or she proposes to make sense of the "data" of linguistic usage. (Consider the various schemata that have been offered by such philosophers as Robert Solomon,[3] Patricia Greenspan,[4] and Robert Gordon.[5]) Furthermore, as a person who not only speaks about emotions, but also experiences them and experiences their connections with actions, perceptions, desires, sensations, and the like, the analyst is also very concerned to make sense of his or her experience and the experiences of other human beings. Thus as I conceive conceptual analysis, it is particularly based on collection of and reflection about examples from everyday human life, many of which can be understood only in the light of a fairly rich narrative background. This preoccupation represents an overlap with literary and psychoanalytic examinations of emotion and a rather strong contrast with biological and neuroscientific examinations. The conceptual analyst, as I understand his *métier*, will look for formulations regarding emotion and particular emotion types, and will be particularly interested in potential counterexamples, also from everyday life, to his formulations.

Conceptual analysis has been criticized as an inadequate approach to the emotions along two different lines by Amélie O. Rorty and Paul E. Griffiths. The two lines of criticism have in common the suggestion that the conceptual scheme provided by our ordinary language about the emotions is a deeply misleading, and perhaps even internally incoherent, indicator of the nature of emotions. Thus any analysis that takes that scheme at face value and as a point of departure is doomed to deep error. Each of these authors promotes an alternative approach. Rorty proposes that we study the history of the philosophies of the emotions because in her view those variegated philosophies have *constituted* the incoherent or apparently incoherent concept of emotion that analysts try vainly to make sense of. Griffiths thinks that the best scientific accounts of the phenomena that we call "emotions" – essentially, accounts from evolutionary biology and its

[3] *The Passions* (Garden City, New York: Doubleday, 1976).
[4] *Emotions and Reasons* (New York: Routledge, 1988).
[5] *The Structure of Emotions* (Cambridge, England: Cambridge University Press, 1987).

auxiliary experimental psychology – show that "emotions" form such a qualitatively diverse set of phenomena that the concept *emotion* and the concepts of particular types of emotion are useless for the purpose of genuine knowledge. I shall examine the arguments and proposals of Rorty and Griffiths, bringing into my critique of Griffiths some observations about the best recent work on the neuroscience of emotions. I shall then end this Introduction with a sketch of a kind of conceptual analysis that avoids the legitimate criticisms that have been leveled against conceptual analysis of the emotions as it was practiced in the 20th century.

1.3. DECONSTRUCTING *EMOTION* VIA THE HISTORY OF PHILOSOPHY

Amélie Rorty begins her paper, "Aristotle on the Metaphysical Status of *Pathe*,"[6] by commenting on the deplorable state of present-day philosophical theorizing about the passions and emotions. The discussions are "arbitrary and factitious" and "puzzlingly pulled in what appear to be opposing directions" (p. 521); these "persistent and unresolvable contemporary polemical debates carry an air of a chimaeral construction" (p. 545). The reason for this apparent impasse is that the concept under discussion itself contains these "opposing directions"; the discussions only reflect tensions internal to the concept:

> We sometimes hold people responsible for their emotions and the actions they perform from them. Yet normal behavior is often explained and excused by the person 'suffering' an emotional condition. We treat emotions as interruptions or deflections of normal behavior, and yet also consider a person pathological if he fails to act or react from a standard range of emotions. Sometimes emotions are classified as a species of evaluative judgments whose analysis will be given in an adequate theory of cognition. But sometimes the cognitive or intentional character of an emotion is treated as dependent on, and ultimately explained by, a physical condition (p. 521).

We can easily think of a few more "opposing directions" that the concept of emotion can pull us in: Some emotions bond people together, others sunder them; some emotions are recognizable via facial expressions, others are not; some emotions disappear as soon as contrary information is heard and believed, others persist in the face of such information; some emotions have an identifiable propositional content, others have none; some emotions (like shame) are intrisically reflexive or self-referring, others (like joy) are not; some emotions are based in the most excellent, others in the most cock-eyed reasoning, while still others are based in no reasoning at all; some emotions are disruptive episodes, relatively unintegrated

[6] *Review of Metaphysics* **38** (1984): 521–546.

into the characteristic concerns and purposes and intentional actions of a person's life, while others are continuous with those leading concerns and express them; some emotions involve discernible bodily arousal, others do not; some emotions are conscious states, others are not; some emotions are pleasurable, others are painful, and perhaps still others are neither the one nor the other; some emotion types are pancultural, others are culture-specific or culture-determined; some emotions are intentional, brought on by the subject for some purpose of her own, while others are not intentional; some emotions are motivations, while others are not. Rorty points out that these "opposing" divisions within the concept of emotion do not tend to be marked by our lexicalized emotion categories ("anger," "nostalgia," "solicitude," "joy"). For example, there might be instances of anger that fall on each of the sides of most of these divisions. Perhaps this fact helps to hide from us the rampant disorder internal to the concept of emotion.

Rorty's thesis that the seeming unresolvability of the debates about the nature of emotions somehow stems from the extraordinary variety and oppositions among the phenomena that we call emotions seems plausible to me, if we add the further premise that the debating theorists base their positions on hasty generalizations from their favored ranges of cases. For example, one kind of theorist fixes on cases of emotion that have highly definite conceptual content, that respond flexibly to changes of information and reasoning, and that are highly integrated into the individual's conscious purposes and explicit worldview. Another kind of theorist fixes on cases of emotion that respond poorly or not at all to information and reasoning, have a strong component of bodily arousal, and have close analogues in beasts and babies. Both theorists then ignore the "opposite" kinds of cases as long as they can, or they authorize their theories by finding clever ways to explain away the counterexamples or assimilate them to their own paradigm, or they just deny that those are "really" emotions. Without the hasty generalizations, followed by digging in of theoretical heels, we would presumably get descriptively richer, less theoretical, monolithic or reductive accounts, ones that would be less controversial because the generalizations would be spare and cautious, always keeping a welcoming lookout for the instructive counterexample. Among people who practiced this more descriptive philosophy of emotion, there would presumably be far less of the unyielding disagreement that Rorty deplores. In making this proposal I am supposing that the concept of emotion is not internally incoherent, and that its apparent incoherence comes from the hasty generalizations of theorists.

But this is not Rorty's proposed resolution of the difficulty. Although she does not go quite so far as to say that the concept itself is incoherent, she does blame the concept at least as much as its analysts. She thinks that our current concept of an emotion is a contraption whose ill-assorted parts are accretions traceable to diverse periods of the history of philosophy in which very divergent agendas shaped the claims that were made about the emotions. If

we lack a clear view of that history, we are doomed to a conceptual muddle, because we take the concept of emotion at "face value"; that is, we treat it as though it is a single, coherent concept.

> The history of discussions of the passions does not form a smooth continuous history, which expands or narrows the class of *pathe* by following a single line of thought. Sometimes the transformations (say from Aristotelian *pathe* to Stoic *passiones*) arise from moral preoccupations concerning voluntary control; sometimes the transformations (say from Renaissance *amor* to Hobbesian passions and desires) are impelled by metaphysical and scientific preoccupations; sometimes the transformations (say from Hobbesian passions and desires to Humean and Rousseauean sentiments) have a political direction. If nothing else, this should show that *pathe*, *passiones*, affects, emotions, and sentiments do not form a natural class. Additions to that class were made on quite distinctive grounds. Before we can evaluate the competing claims of current polemical debates, before we can understand the force of their various claims, we must first trace the philosophic preoccupations in which they originated (p. 545).

Again, it is not entirely clear whether Rorty is claiming that, for example, the Stoics merely noticed and emphasized that some emotions are subject to voluntary control and had a theory about it and built further theory on it, perhaps overgeneralizing from it, or whether the Stoics *invented* voluntary control of emotions and then passed that trait of emotions (or at least of the concept of emotion) on to us. If the former is so, then it might be *interesting* to know what the Stoics said about voluntary control, but it would not be *necessary* for a contemporary conceptual analyst, as Rorty seems to suggest it is. The analyst would be looking at an emotion like anger and noticing the same feature that the Stoics exploited, namely that people can often control their anger if they have a modicum of understanding of their emotion and make some effort. Since the conceptual analyst would be doing essentially the same kind of thing the Stoic was doing, the analyst would be under no necessity to advert to what the earlier theorist had said.

If the present-day theorist really needs to know the Stoic discussion, the latter must be somehow constitutive of the very subject matter of the present discussion. In that case when Rorty says, "Before we can evaluate the competing claims of current polemical debates, . . . we must first trace the philosophic preoccupations in which they originated," she must be saying that the "opposing" features that set the parameters of our debate actually *originated* in the earlier philosophical discussions. For example, if we can't appreciate the notion that emotions are subject to voluntary control without knowing the Stoic contribution to the subject, then the fact that emotions are subject to voluntary control is not just *noticed* by the Stoics but *created* by them. Even this would not be enough, strictly speaking, to make acquaintance with historical Stoicism a necessary condition for understanding current debates because the voluntariness of emotions might take on a life of its

own after having been socially constructed in terms of Stoic theory. On this interpretation, Rorty's claim that we cannot understand emotions without history of philosophy implies that this history not only created such features of emotions as their voluntariness, reliance on judgments, power to deflect normal behavior, grounding in physiological conditions, and so on, but created these features in such a way that they are *internally tied to the originating theories.*

What kind of understanding of the concept of emotion would emerge from a study of the history of the philosophy of emotion, on the second interpretation of Rorty's thesis? Since by hypothesis our concept of emotion is socially constructed in such a way as to make conceptual–analytical accounts of it chimaeral, the result of the historical studies that Rorty envisages would be our understanding of an incoherent "concept" *as incoherent.* If we wrote the history of the concept of emotion, we would understand *emotion* to be a philosophically constructed chimaera (my dictionary says a chimera is "an imaginary monster compounded of incongruous parts"). We would see that the concept of emotion has no real referent, but only this constructed, chimaeral one. This history would explode a myth, exposing a *purported* concept for the monstrous contraption that it is.

We might wonder why, on this interpretation, the unmasking of the "concept" of emotion could not proceed ahistorically, just by showing the internal contradictions in the concept. Perhaps the idea is that this procedure would never decisively show the concept to be incoherent since a conservative could always fall back on the hope of a future account that will show the concept's coherence. The genealogy of *emotion* might be thought capable of laying this hope finally to rest, by showing once and for all where the contradictory strands in the "concept" came from.

It is not clear to me that Rorty endorses the rather implausible view that I have just sketched. Perhaps she thinks that the influence of philosophical theories on our concept of emotion is of some looser variety, and that phrases like "must first trace the philosophical preoccupations" and "necessary to trace the history" should be taken more weakly than I have done. She does make one remark that seems to make the history of philosophy less crucial:

> Officially we are preoccupied with determining whether emotions can be evaluated for their rationality; or whether they are voluntary; or whether they can be "reduced" to cognitions; or whether they are interruptions of behavior that is normally purposeful. But in fact we know better: when we are really thinking, rather than making pronouncements, we know that we evaluate the appropriateness of emotions by criteria that are much richer than those of logical consistency: we are interested in determining whether they are inadequate or excessive, crude or subtle; whether they are harmoniously balanced with one another; whether we admire the character traits they reveal and the motives that usually accompany them. And when we are careful, we usually also distinguish passions, emotions, affects, sentiments (pp. 521–522).

While I would not describe in just Rorty's terms the kind of conceptual analysis I commend, I agree with the direction of her thought in this quotation. She is saying, in effect, that if we stop crudely theorizing and look carefully at the human emotions and our modes of describing and evaluating them, if we stop thinking in terms of simplistic questions about emotions and *look* to see how they actually and richly function in the course of our lives, then the seeming incoherence in the concept of emotion begins to disappear and we see not incoherence and in principle irresolvable debates, but subtle and rich variety linked by family resemblances. So perhaps Rorty is admitting that we may not strictly *need* the history of philosophy after all, but just a more astute and careful and "empirical" and less theoretically hidebound application of philosophical analysis. But because philosophers have historically picked up on some features of emotions to the exclusion of others, the history of philosophy might help in our analysis by alerting us to features that need accommodating and abstractions we need to avoid. On this interpretation, which we might call the "history of philosophy as aid to conceptual analysis" view, Rorty would not be saying that the concept of emotion is an imaginary monster, nor that the history of philosophy is strictly necessary to its analysis. The history of the philosophy of emotions is *a useful but non-necessary adjunct* to philosophical analysis (along with several other adjunct disciplines), in heading off theoretical dead-ends, raising interesting questions, and making interesting proposals.

My purpose is not to adjudicate the interpretation of Rorty's provocative paper, but to defend a kind of conceptual analysis of the emotions. Since the second interpretation allows for conceptual analysis with a recommendation of aid from the history of philosophy, I have no quarrel with it. And I am interested in the first interpretation, not because I ascribe it with confidence to Rorty, but because it is a challenge to my project.

Let us try out an argument, which we might call the realist common sense objection, against the historically constructed chimaera theory (HCCT). As a proposal for examination why not say the following:

Proposal:
We can explain the "opposing" features of emotions much more straightforwardly. We needn't posit that the history of philosophy has created these features, because we can observe them in our everyday experience. For example, we can explain why people have thought that emotions are strongly connected with judgments by noting that people, in any historical period, including our own, can be roused to anger or fear or nostalgia by narratives, and that their anger or fear can often be dispelled instantaneously by telling them something. We needn't resort to the history of philosophy to explain why people think emotions are grounded in physical conditions such as fatigue or the influence of drugs; appeal to their experience is enough. We do not need the history of philosophy to explain why people are sometimes held responsible for their emotions and sometimes exonerated because of them. Nor do we need it to show us why people think that both normal

and abnormal human functioning depend on emotional states and dispositions. These judgments about emotion can be nicely attributed to the human experience of living. And clearly, the philosophers who built their theories on one or another of these features did so by observing them, just as we do. HCCT reverses the order of priority: the philosophers' theories came from the features, not the features from the theories. And if our attributions of these seemingly opposed features to emotions are results of observation rather than of theory construction, then we may have some confidence that they only seem opposed – that the concept of emotion is not a chimaera but a consistent body of attributions. After all, reality, even psychological reality, is not likely to be incoherent.

The weakness of this response is that on HCCT, the fact that we can observe the opposing features is not evidence that they were not created (in the strong sense required by HCCT) by the history of philosophy. As Rorty says, "All these views are embedded in our common speech and common sense, as well as in the literary works that form our understanding of ourselves" (p. 545). So the position is insulated against the common-sense realist objection. But HCCT needs to have more going for it, if we are to abandon common-sense realism for it, than that it is insulated against objections from common-sense realism. We need some positive reason to accept it, since common-sense realism is common sense. If philosophical reference to each of the features of passions that Rorty finds identified and exploited in the history of philosophy can be as well accounted for on the hypothesis that the philosopher in question identifies a previously existent feature as on the hypothesis that the philosopher invents the feature and then passes it down to us in the form of common sense, then the history of philosophy gives us no reason to accept HCCT rather than common-sense realism about emotion features. In that case we just have an evidential stand-off; and since common sense takes natural precedence, we have no reason to abandon it.

But other considerations seem to weaken further the appeal of HCCT. We might wonder where philosophers got the idea of the feature – say, that emotions are dependent on judgments or that emotions disrupt normal behavior or that emotions are necessary to fully normal behavior – if they did not get it from observation. Philosophers are typically pretty creative people, and so we might think there's no mystery here, but my guess is that if we looked at the contexts in their writings in which philosophers identify the features that have come to play roles in modern discussions of the emotions, we would see that they often appeal to examples and observations. This is certainly true of Aristotle and Hume, and I would guess that it is true of most of the main players in Rorty's history of the emotions.

Also, we might wonder why these features have had such sticking power in human life and why they are sustainable at pretty much all educational levels and with so little direct influence from the history of philosophy. We might think that where concepts are invented more or less out of whole cloth and without much of an observational basis, they require more direct

and continuous intervention from theorists than the concept of emotion seems to enjoy. Another possible argument might be launched by examining anthropologists' studies of emotions among peoples who cannot have been influenced by the history of philosophy. If such studies show the natives identifying features of emotions like the ones that generate the controversies in recent Western intellectual discussions, that would be evidence that these are observable features antedating philosophical theories that exploit them (see Sections 3.2b, 3.3b, and 3.3c).

Yet another potential argument is that if we expand the list of "opposing" features, as I did at the beginning of this section, we may begin to have a hard time finding plausible originating points for them in the history of philosophy. We may wonder why Rorty selects just four oppositions, and whether all four even of these are plausibly explained in terms of the history of philosophy. In any case, the project of showing that the "opposing" features of emotions were born in the history of philosophical discussions of the emotions has yet to be done. The hypothesis cannot be fully evaluated in the absence of a more or less full, book-length demonstration.

Let us consider the history of philosophy as an aid to conceptual analysis. Rather than think of the history of the philosophy of emotions as *constituting* or creating our concepts and experiences of emotions, we might think of this history as *influencing* them, in the course of responding to the phenomena. Different players in that history respond according to their own particular agendas and theoretical frameworks, so that they highlight different features of the emotions, which, as we have seen, do have many diverse features. The anthropologist Robert Levy has proposed that societies may "hypercognize" or "hypocognize" emotion types. For example, the Tahitians, among whom Levy did field work, hypercognize anger but hypocognize sadness. They have a subtle vocabulary for describing, explaining, evaluating, and prescribing for anger but not even a word that denotes sadness. The Tahitians do become sad, but they are less likely to notice it and do not identify it with the same precision as societies in which it is more "cognized."[7] Something similar might be true of the generic features of emotions: For theorists with differing interests, different features will be salient, and the saliencies will both influence and result from their theories; but this is not to say the features are created by the theories. Perhaps Aristotle, the Stoics, Augustine, and others did not create the features that our concept of an emotion attribute to emotions; instead, they all more or less successfully describe phenomena that have been relatively stable through human history, the same kind of thing that contemporary analytical philosophers, anthropologists, psychologists, neuroscientists, and evolutionary biologists are giving

[7] See "Emotion, Knowing, and Culture," in Richard A. Shweder and Robert A. LeVine (eds.), *Culture Theory: Essays on Mind, Self, and Emotion* (Cambridge, England: Cambridge University Press, 1984), pp. 214–237.

their accounts of. Emotions invite highly perspectival accounts because they are many-sided phenomena. On this picture we may admit possible influences from Aristotle, the Stoics, and so on, on our way of thinking about the emotions, but it would be an exaggeration to talk about the transformation of Aristotelian *pathe* into Stoic *passiones*, as though the subject matter of their discussions is not the rather old familiar facts of anger, fear, joy, and hope. Instead, we could talk about Aristotelian *ideas* about emotions and Stoic *ideas*. In that case the puzzles we experience when we study the emotions as philosophers would be not just products of this history but, more importantly, products of the phenomena – the emotions that we observe in human beings. And the supposed conflicts that we find within the Aristotelian account, or the conflicts between that account and, say, the Stoic account, would be due as much to the actual features of emotions and passions as to theorists' accounts of them. An imperialistically social constructivist account of emotions is as far from the truth about them as a purely neurological account. Each, according to its special interests, "hypercognizes" certain features.

Let us distinguish emotion category concepts from emotion type concepts. Examples of category concepts are ones that have roughly the same degree of generality as *emotion: sentiment,* πάθος, *passio, affect, affectus, passion*, and so on. Type concepts are concepts of subclasses within the categories of emotion, passion, affect, and so on. Examples are *anger, dismay, sorrow, shame,* τὸ νεμεσᾶν,[8] *liget,*[9] and so on. Emotion category concepts encompass a range of type concepts. Thus *emotion* is the class that includes anger, nostalgia, shame, joy, and perhaps (on the periphery) puzzlement, amusement (at humor), surprise, and the startle response. Because *emotion* and *passio* and πάθος encompass partially different ranges of types, they will be different, though largely overlapping, concepts. For example, Thomas Aquinas lists desire (*concupiscentia*) as a type of *passio*, whereas we would probably not regard it as a kind of emotion (though we might include it among the passions); and as Rorty makes abundantly clear in the main body of her paper, the concept of a πάθος in Aristotle's society was much broader than our concept of emotion, encompassing such things as bodily wounds and states of sense perception. I think that the studies in the history of philosophy that Rorty commends can sensitize us to the variability of the category concepts related to that of emotions (passions, sentiments, etc.) and to the relativity of such variation to human interests; they can mitigate a certain platonizing tendency in the study of emotions, a tendency that natural languages

[8] An emotion type discussed by Aristotle, different from envy (φθόνος), characterized by discomfort about someone else's undeserved good fortune. See *Rhetoric*, Book II, Chapter 9 (pp. 1386b10–1387b20).

[9] A dominant emotion in the moral psychology of the Ilongots, a head-hunting group in the Philippines (see Section 3.3b).

seem to engender. The history of philosophy and psychology is full of lists of "basic" emotions, and these lists differ remarkably from one another; the best explanation of this diversity seems to be that the lists reflect different sets of theoretical interests (see Section 3.1c). Also, the history of philosophy, like cultural anthropology, can moderate our naive tendency to think that our emotion type vocabulary divides the world of the emotions in the natural and only possible way (see Section 1.5e).

So the concept of emotion can be thought of as determined by the range of type concepts that it encompasses, but it must be admitted to be somewhat indeterminate because of questionable types on the outer fringes, such as surprise, startle, amusement (e.g., at jokes), interest (e.g., in philosophical ideas), and others. The intuitions of good speakers of English vary as to whether these states are emotions. But the bare question of English usage is not in itself a very interesting one; we want to know *why* type concepts like *anger*, *fear*, and *envy* are solidly in everybody's paradigm of emotion while *surprise* and *startle* are only in some people's. One way to get at an answer to this question will be to take seriously the various "opposing" features of the paradigm cases that Rorty's essay invites us to highlight, as well as others that I have indicated. If we can come up with a broad unifying conception of emotion that accommodates all these opposing features, then we will have a conception that unifies at least the paradigm cases and gives us a plausible account of why English speakers group this range of mental states together under a single class name. That is the main task of Chapter 2.

In Chapter 3 I will then test the conception by analyzing a wide range of type concepts, including not only the paradigm types but pretty much anything that anybody is inclined to call an emotion, including the contested types. My strategy will be not so much to try to decide whether each type belongs to the category of emotion as to try to see in what ways each type is similar, and in what ways dissimilar, to the undisputed paradigm cases. Thus I do not offer my account as a "theory," as implying that all and only what we would properly call an emotion fits the proffered conception. Instead I shall argue that the conception is superior to its competitors in making sense of all the "opposing" features in the paradigm cases. I shall try to show fairly precisely the various ways in which the other cases deviate from the paradigm. But despite the fuzziness on the edges, I think I will have shown that the concept of emotion is not a monster.

1.4. DECONSTRUCTING *EMOTION* VIA THE LIFE SCIENCES

There is a strong movement these days to subsume psychology under biology and related disciplines such as physiology and especially neurophysiology. Paul Griffiths's *What Emotions Really Are: The Problem of Psychological*

Categories,[10] is an especially explicit, philosophically sophisticated, and uncompromising example of this trend. Besides this, the book is focused on emotions and directly attacks conceptual analysis as an approach to understanding emotions. For all these reasons, it is interesting for our purposes, and I hope I will be forgiven for paying so much attention to it.

a. Science Fractures a Concept?

Griffiths's thesis is reminiscent of Rorty's proposal that we deconstruct the concept of emotion using the history of philosophy because, trading on some of the "oppositions" that are present in the ordinary concept of emotion, he proposes that under scientific study the concept of emotion will "fracture" into three radically distinct concepts. Recent science shows that the vernacular concept of emotion covers a range of things that have as little to recommend their assimilation under a single concept as the hodgepodge in Aristotle's class of superlunary objects.

> Emotion is like the category of "superlunary" objects in ancient astronomy. There is a well-defined category of "everything outside the orbit of the moon" but it turns out that superlunary objects do not have something specially in common that distinguishes them from other arbitrary collections of objects. . . . what we know about ["emotions"] suggests that there is no rich collection of generalizations about this range of phenomena that distinguishes them from other psychological phenomena. They do not constitute a single object of knowledge (p. 14).

In particular, he is impressed by the "opposition" between (a) emotions that show a clear physiological syndrome, are reflexlike, pancultural, and phylogenetically ancient, and do not require higher cognitive processing and (b) emotions that do require such processing and may be quite culturally specific and do not show any clear physiology.

> The first group he calls (following Paul Ekman) "affect program responses." . . . the affect program theory deals with a range of emotions corresponding very roughly to the occurrent instances of the English terms "surprise," "fear," "anger," "disgust," "contempt," "sadness," and "joy." The affect programs are short-term, stereotypical responses involving facial expression, autonomic nervous system arousal, and other elements. The same patterns of response occur in all cultures and homologues are found in related species. These patterns are triggered by a cognitive system which is "modular" in the sense that it does not freely exchange information with other cognitive processes (p. 8).

"Higher cognitive emotions" are divided into two discrete categories. Griffiths calls the first category "irruptive motivations," following Robert Frank.[11] These are like the affect program responses in that both kinds of

[10] Chicago: University of Chicago Press, 1997.

[11] *Passions Within Reason: The Strategic Role of the Emotions* (New York: Norton, 1988).

emotion "produce a form of passivity" (p. 245); that is, they are not intentionally produced but *come over* the subject in response to something. These states, which include instances of loyalty, jealousy, and guilt, as well as episodes that vernacular speech would identify with the same names as are used for the affect programs, are "states which interfere with the smooth unfolding of plans designed to secure our long-term goals" (p. 246). Thus they are not only *irruptive* (i.e., passive states) but also *disruptive* of long-term goal-seeking. An example of irruptive motivational anger would be the emotion of a man that drives him to take revenge on people for trespassing his rights even when taking revenge undermines his considered long-term goals (e.g., making money, keeping his friends). Frank argues that such an emotion is evolutionarily adaptive, despite first appearances, because people will be disinclined to trespass the rights of a person who is likely to go ballistic in this way. The irruptive motivations have surface irrationality that hides a deeper function. These "emotions" may occur in the absence of facial expression and autonomic arousal and do involve higher cognitive processing. In our example, the concept of a violated right, which the angry subject deploys in his response to the situation, clearly requires the functioning of "higher" parts of the brain, not just the "informationally encapsulated" ones that operate in the affect programs.

The second kind of higher cognitive "emotions" are "socially constructed." Griffiths distinguishes two kinds of social construction, the "social concept model" and the "social role model," and dismisses the former as trivial. He points out that many social constructionists in emotion theory think that a society constructs emotions by providing categories in terms of which its people respond emotionally to objects and situations. But this "is a model of the emotions themselves only because an emotion is identified with the thought that the eliciting situation is present" (p. 139; here Griffiths refers to a version of the propositional attitude theory that we will discuss in the next subsection). The kind of socially constructed emotions that fill a significant category are the ones he calls *disclaimed actions*. These are essentially fake emotions – behavioral patterns that one produces, under the guidance of cultural rules, for the sake of achieving some goal. Thus, according to Griffiths, they lack the "passivity" that he finds common to the affect programs and the irruptive motivations. Far from disrupting goal-directed behavior, these are stratagems to purpose. Griffiths hastens to point out that the subject of such an "emotion" is not merely pretending: "The subject does not have conscious access to the causes of their [*sic*] behavior and provides an erroneous explanation of their behavior that masquerades as an introspective report" (p. 158). Borrowing from Robert Solomon, he says,

> A good example is the display of anger as an unconsciously implemented "strategic behavior" in a marital quarrel. The agent has reasoned that they can improve their

position by adopting the role of someone who believes themselves wronged. The agent is not simply acting, because although they are motivated by these considerations they are not aware of this motivation. In such a case the agent will behave as if they had judged themselves to be wronged. There is all the difference in the world between this and actually believing oneself to have been wronged (p. 233).

These "emotions" are so different from the items that belong in his other two categories that Griffiths has some doubts about whether to include this class as one of the kinds of things into which the vernacular concept of emotion fractures:

The disclaimed action emotions cannot be placed in a single category with the other emotions because they are essentially pretenses. It would be like putting ghost possession in the category of parasitic diseases. Averill and Boothroyd (1977) suggest that "falling in love" is the adoption of a social role which licenses the performance of certain behaviors. If so, then just as there are no ghosts to explain ghost possession, there is no state of love to explain love behavior (p. 246).

But since Griffiths is sometimes willing to countenance their inclusion, let us say that, according to him, what folks call emotion, when subjected to scientific pressure, fractures into three concepts: affect program responses, irruptive motivations, and disclaimed actions.

We will want to ask two questions about this schema. First, what is the scientific rationale for it, and how does that rationale impinge on the project of giving an account of the emotions for purposes of a psychology of the virtues? And second, are the proposed categories really discrete? Do they "carve nature at its joints"? Are they natural kinds? Do they neatly divide the area that folks call "emotions" and provide a replacement schema that makes more sense of the phenomena? But before we turn to these questions, we must look at Griffiths's direct critique of conceptual analysis as an approach to understanding the emotions.

b. Conceptual Analysis as Propositional Attitude Analysis

Griffiths thinks that conceptual analysis is strongly, if not essentially, tied to what he calls the "propositional attitude" theory of emotions. According to the variants of this theory, emotions are either simply evaluative beliefs, or evaluative beliefs plus some added feature such as physiological perturbation, desire, or "affect." Thus a person's being in the state of fear is equivalent to her believing she is in danger, or to this belief plus an appropriate kind of physiological arousal, or the desire not to be in danger, or a certain phenomenological tone called "affect." Griffiths raises a number of objections to the propositional attitude theories, such as that a subject's beliefs often conflict with his emotions (he believes earthworms to be harmless but still fears them), that people can have the relevant evaluative beliefs yet feel no emotion (I believe I have sinned, but I don't feel guilty), that people

respond with emotion to presentations of what they know to be fiction (e.g., in novels and movies), and so on. It is true that conceptual analysis of emotions has been dominated, since the 1960s, by propositional attitude theories, and these objections and others devastate the propositional attitude theories. But the objections require nothing beyond the resources of conceptual analysis: a rich enough fund of examples from ordinary experience and some careful reflection. I have previously raised these objections and others[12] and will discuss them in Section 2.4.

A somewhat more significant objection to conceptual analysis is that it "can tell us only what people currently believe about emotion" (p. 39), not about what emotions really are, in a sense that is determined by their underlying causal mechanisms. This strong association of what emotions really are with a certain kind of underlying causal mechanism is the crux of Griffiths's argument against conceptual analysis. The kind of causal mechanism he has in mind is the kind that evolutionary biology and neuroscience, as branches of physical science, try to establish. Let us say that some state or syndrome of states that we could call "fear" in human beings can be shown to be present (with variations) in all animal species that bear a certain evolutionary proximity to us (say, the primates, or, less proximately, the mammals). This would yield a causal explanation of human fear in terms of the mechanism of *descent*: Humans have this syndrome because their ancestors had something like it. That humans' ancestors had something like it and that we have retained the trait can perhaps be explained by a second explanatory strategy, *adaptation*: Because fear involves avoidance of and escape from dangers, fear seems to promote survival. Associated with these kinds of explanation would be the expectation of cross-species neurological similarities in the fear response. For example, we would expect to find significant parallels in the neurological circuitry and chemistry of fear in rhesus monkeys and human beings and somewhat weaker but still significant parallels in rats. Explanations of fear in terms of the causal mechanisms of *neural circuitry and chemistry* would thus be another kind of explanation that, in Griffiths's view, would count as contributing to our knowledge of what an emotion "really is." Since these patterns of explanation require stable and determinately structured causal mechanisms ("causal homeostatic mechanisms") and since such mechanisms determine natural kinds, Griffiths is preoccupied with establishing natural kinds as a basis for explanation in psychology. The fear affect program is such a natural kind, as is, presumably, irruptive motivational fear (though Griffiths admits that we have little solid information about what the homeostatic causal mechanism is in this case). Clearly, if what we mean by "what an emotion really is" is the homeostatic physical

[12] "Solomon on the Control of Emotions." *Philosophy and Phenomenological Research* **44** (1984): 395–403; "What an Emotion Is: A Sketch." *The Philosophical Review* **97** (1988): 183–209; "Propositions and Animal Emotion." *Philosophy* **71** (1996): 147–156.

mechanism that underlies the emotion, then we will not get very close to an answer by asking what ordinary people, in the ordinary course of their ordinary experience, have come to believe about emotions.

On the other hand, if What is an emotion? is not a question about underlying physical mechanism (whether thought of in terms of distant history or present operation), but about emotions as experienced by human subjects, as structures of meaning and explanation in the course of social life, as entering into our actions and reasoning, as evaluated to be proper or improper, praiseworthy, blameworthy, or morally indifferent, and as bearing on our happiness and maturity and relationships with one other and with God (see the passage from Trollope with which this chapter begins), then conceptual analysis may be the central approach to determining "what an emotion really is."

According to Griffiths, conceptual analysis of fear is just a description of how people in the analyst's society use the word or what they believe about the word's supposed referent. Something analogous but presumably more general would be true of conceptual analysis of "emotion." Analysts of this ilk dismiss empirical findings about fear, such as Paul Ekman's evidence that facial expressions of fear are pancultural and interculturally recognizable and Joseph LeDoux's discoveries about the neural circuitry involved in the production of fear. They regard this information as irrelevant because it is not known to most users of the word "fear" and therefore cannot be part of the meaning of "fear." They are, according to Griffiths, not *au courant* of the philosophy of language which, since the work of Saul Kripke and Hilary Putnam, sees that linguistic usage can change in response to scientific discoveries. For example, "fish" once included whales, but the discovery that whales are mammals has changed that, at least for sophisticated speakers.

> If fear is a putative natural kind like water or crustacean then a causal theory [of meaning] would say that "fear" has as part of its meaning a schema awaiting the results of future research. Fear is "whatever is happening to people in these paradigm cases" (pp. 4–5).

If it turns out that "fear" is ambiguous as between more than one naturally distinct category of states as defined by their homeostatic causal mechanisms, then science might lead to a revision of the extension of "fear," in much the way that it has led us to exclude whales from the extension of "fish." For example, if affect program fear is as different from disclaimed action fear as Griffiths thinks it is, biology and neuroscience might lead us to stop thinking of disclaimed action fear as a sort of fear. We might instead adopt the policy of calling it "fake danger response."

In the following sections I shall argue that it is implausible to suppose that the field of things that folks call emotions can be divided, for scientific purposes, into the three categories that Griffiths has proposed. The phenomena interlock in ways that make these categories very awkward and artificial. To

put the matter in Griffiths's terms, affect programs, irruptive motivations, and disclaimed actions are not natural kinds. Or, to put the matter in Rorty's terms, the "oppositions" of attribute among the things that we call human emotions do not divide them into the kind of neatly discrete categories that Griffiths's theory needs them to divide into. His categories might be useful for certain restricted purposes; for example, one might undertake a study of just those things that fit the category of irruptive motivations, simply because they fit the pattern of explanation that Robert Frank proposes for them. But to do so would be to leave out a number of other kinds of "higher cognitive" emotions that do not fit the category, including some that we might call "higher cognitive affect programs."

I shall argue that, given the ways that the various "opposite" attributes of emotions criss-cross the whole field, defeating any neat categorial scheme, it is better to think of what folks call emotions as not fracturing into several natural kinds, but as belonging to one category, albeit one that is fuzzy on the edges and held together in part by family resemblances and in part by a "homeostatic causal mechanism" of a kind rather different from what Griffiths has in mind. I shall argue that while studies of emotion by physical science, along the lines commended by Griffiths and LeDoux, are possible for certain ranges of emotions (especially the affect programs) and will no doubt yield interesting results, they are not likely to account for the whole range of emotions as well as the best conceptual analysis, supplemented with scientific knowledge, can do. And especially for purposes of moral psychology of the virtues and vices, where the higher cognitive shaping of the emotions (including the affect programs) is so important, conceptual analysis is the indispensable central method.

c. Affect Programs

The affect program responses are clearly the most encouraging kind of emotion for the biologically oriented psychologist because they yield more readily to biological explanatory strategies than do the items that fit in Griffiths's other two categories or other emotions that fit in none of these categories. Paul Ekman has spent more than thirty years studying responses that he calls "anger," "fear," "sadness," "enjoyment," "disgust," and "surprise," and he has gathered empirical evidence that the "emotions" he studies are similar to one another but distinct from other kinds of psychological states (e.g., beliefs) in having characteristic facial expressions that can be cross-culturally recognized with fairly high reliability and distinctive autonomic arousal, among other marks; he has also gathered empirical evidence that these states differ from one another by the *particular* facial expressions and *particular* patterns of autonomic arousal that are characteristic of them. Griffiths adds that these emotions are "informationally encapsulated" – that is, controlled by a kind of information processing that is very rapid and

largely unconscious and not susceptible to direct modification by higher cognitive processes. This is one of the chief ways in which Griffiths keeps the affect program responses distinct in kind from the "higher cognitive" emotions. Griffiths and Ekman both stress that states strongly homologous to these six emotions can be found in species evolutionarily close to *homo sapiens*.

Joseph LeDoux's neuroscientific work on fear[13] is a good example of the kind of emotion research that Griffiths commends. A chief research strategy used to locate brain functions is to destroy, by highly selective surgery, some part of the brain and then to observe which functions have been lost. Another technique is to inject a kind of stain into one part of the brain and scare the subject, then to kill the subject and look at other parts of the brain under a microscope to see where the stain got projected through the neural network. Because humans have been slow to volunteer as subjects in this kind of research, much of it is done on animals such as monkeys, rabbits, rats, and pigeons, who are not consulted. We have seen that Griffiths stresses the importance of homologies, and not just of adaptive strategies, in explanations of emotion, and he ties homologies to homeostatic causal mechanisms. Species evolutionarily distant from human beings, such as wasps and clams, have response patterns that are analogous to fear inasmuch as they are responses to potential harm and involve avoidance of the harm. It even turns out that a surprisingly wide range of animals (including the fruit fly and marine snails) can be conditioned to respond with fear to new stimuli (p. 147). Presumably these responses, like human affect program fear, are an adaptation to dangerous environments, and there is some kind of causal mechanism by which the response operates. But wasps are so unlike human beings that it would not be a very promising research strategy to correlate surgical ablations of wasp counterparts of brains with the resulting functional deficits, if the purpose was to discover the causal mechanism underlying *human* fear. Wasp fear and human fear are only analogous, while monkey fear and human fear are homologous: because monkeys and people are evolutionarily closely related, their fears (at least a certain range of what we call fears) have very similar underlying causal mechanisms. Where we are not sure of the extent of similarity between the two mechanisms, the best bet, Griffiths is saying, will be with organisms that are known to be closer in descent.

Let me give you an example of the kind of explanation LeDoux's work yields. Wanting to discover the brain mechanisms underlying fear responses to a conditioned auditory stimulus, he applied the surgical and staining techniques I mentioned above to rats that he conditioned (or tried to condition, after surgery that sometimes prevented conditioning) to a sound by pairing it with an electrical shock. A normal rat, so conditioned, would respond to

[13] *The Emotional Brain: The Mysterious Underpinnings of Emotional Life* (New York: Simon and Schuster, 1998).

the sound, in the absence of the electrical shock, with the marks of affect program fear: muscular "freezing," increased blood pressure and heart rate, reduced pain responsivity, and elevated stress hormones from the pituitary gland (p. 144). To find out how the normal rat's brain works, LeDoux systematically damaged various brain parts, starting with the "highest" part of the auditory pathway. At each stage he would attempt to condition the damaged rat to see whether it could still be conditioned and, if so, whether any parts of the fear response were missing. He started by damaging the auditory cortex, and he found that this damage had no effect on the rat's conditionability. So he then damaged "the next lower station, the auditory thalamus, and these lesions completely prevented fear conditioning" (p. 152). From this combination of effects he inferred that the auditory stimulus had to pass through the thalamus but could bypass the "higher," more discriminating, auditory cortex. Where did it go from the auditory thalamus? To find out, he injected stain into the rat's auditory thalamus, scared the rat by using the conditioned stimulus, and put slices of the rat's brain under the microscope. The stain he had injected into the thalamus could be seen in four different subcortical regions, so one of these areas was probably where the responses characteristic of fear are generated. A process of further elimination was needed to determine which one, so LeDoux damaged each of these parts in some more rats' brains, to see which kind of damage prevented fear conditioning. He found that damage to the amygdala was the kind that prevented such conditioning. More particularly, damage to the central nucleus of the amygdala interfered with all the different measures of conditioned fear, but he also discovered that each of these responses is brought about by different outputs of the central nucleus.

> For example, I demonstrated that lesions of different projections of the central nucleus separately interfered with freezing and blood pressure conditioned responses – lesions of one of the projections (the periaqueductal gray) interfered with freezing but not blood pressure response, whereas lesions of another (the lateral hypothalamus) interfered with the blood pressure but not the freezing response (pp. 158–159).

The circuitry identified for conditioned fear by the investigations I have summarized is thus something like the following: (1) A conditioned stimulus activates neurons in the ear, which project to (2) the auditory cortex and (3) the auditory thalamus. The auditory thalamus sends signals to (4) the central nucleus of the amygdala, which then projects to (5, etc.) several other parts of the brain which, no doubt by yet other mediations, bring about the bodily and behavioral marks of fear. This is of course only a crude and partial account of the causal mechanism underlying fear. Neuroscientists have much to say about the mechanics of the transmission of signals along the neural pathways, and we have said nothing about the feeling of fear (if rats do feel fear), which involves the complicated and controversial neuroscience of consciousness. But I think I have reported enough to indicate

the *kind* of explanation of emotions that biologically oriented psychologists look for. They are explanations in terms of physical underlying causal mechanisms.

Clearly, LeDoux's scientific work is limited to the emotions that Ekman calls affect program responses. The importance of the sensory thalamus and relatively less importance of the sensory cortex and complete absence of essential involvement of language centers in LeDoux's fear reflects what Griffiths calls its "informationally encapsulated" nature. The fact that conditioning is the only strategy for emotional learning that is in view for LeDoux's research is another indicator that he is dealing exclusively with "lower" cognitive processes. The fact that LeDoux can conduct his research almost entirely by manipulating and observing rats reflects the evolutionary homology that Griffiths attributes to the affect programs. And it is equally clear that any emotion that lacked clear and distinctive physiological markers would be quite foreign to LeDoux's research program.

But we can here raise a question that Griffiths's metaphysics discourages us from asking: Are the boundaries of the category of affect programs set by the nature of the phenomena under investigation, or by the limitations of scientific technique and interests? It is one of Griffiths's main theses that natural kinds are the posits of our best scientific theories. But this will not be so if science ignores certain cases simply because they do not fit the theory or are not susceptible to currently available research techniques. In that case it may be better to admit that current science cannot explain all the cases that seem to belong together.

LeDoux does not deny the possibility that some instances of adult human fear may depend on some of the highest cognitive functions. In other words, he does not hold that the fear that he researches is "informationally encapsulated," though he does show that fear *can* occur without the involvement of the "higher" brain centers. But his research does not purport to give a full neural explanation for instances of fear that do involve "higher" brain centers. Although he does not speak of natural kinds, it is as though he allows that there may be instances of the natural kind fear that his neurological theory of fear does not explain. Antonio Damasio[14] is even more explicit in denying that what Griffiths calls affect program fear is necessarily informationally encapsulated.

We have seen that one of the defining characteristics of affect program states as a supposed "natural kind" is that they are "informationally encapsulated" and that they thus belong in an entirely separate category from the "higher cognitive emotions." Othello's sexual jealousy cannot be an affect program, because it is mediated by Othello's understanding a narrative or implied narrative. For his jealousy to be an affect program, "he would

[14] *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Avon Books, 1994).

have had to catch Desdemona in bed with Cassio, or at least have seen the handkerchief, before his jealousy was initiated" (Griffiths, p. 117). As Griffiths says at the end of the book, "no one expects discoveries about the fear affect program to apply to responses to danger mediated by higher cognition" (p. 242), speaking presumably of someone who has been convinced by the argument of his book. (Calling both these kinds of things "emotions" is on a par with calling the sun and some Martian fossilized bacteria "superlunary objects.")

But if natural kind concepts are supposed to carve nature at its joints, these do not seem to be natural kinds because many cases of emotions span the categories. Consider the following case. Just prior to lunch time, a hungry philosopher, browsing in the library, comes on a review of his recent book in a prominent journal and, as he reads, he realizes that the reviewer has found a fatal objection to his central argument, an argument on which he has built his career, his reputation, and his sense of his own professional worth. He breaks out in a cold sweat, his heart starts pounding and his blood pressure goes up, the muscles in his back and neck tense up, his mouth goes dry, and his appetite disappears. His involuntary facial expression is one of extreme distress. This appears to be an affect program emotion mediated by higher cognitive processes. It looks for all the world as though the information from his cerebral cortex (the philosopher is reading sophisticated theoretical material and understanding both its theoretical significance and the significance it has for his career, his reputation, and his sense of importance) is getting through very nicely to his amygdala and having quite an impact on it. Furthermore, the philosopher is sophisticated enough to realize that he is in a state of panic, and exactly why. Here, then, is an affect program state that appears not to be "informationally encapsulated." Furthermore, the affect program is not "triggered" in a reflexlike way as Griffiths's affect programs are supposed to be triggered but comes on gradually as the realization dawns on our philosopher that the jig is up for his theory. It is an affect program state that is also a higher cognitive emotion.

It is just this kind of case that Antonio Damasio is trying to explain when he claims that the "limbic system" (the more "primitive" and evolutionarily older part of the brain in which much of emotional processing is supposed to go on, and includes the amygdala) is accessible to the cerebral cortex (see Damasio, pp. 131–139):

> In many circumstances of our life as social beings . . . we know that our emotions are triggered only after an evaluative, voluntary, nonautomatic mental process. Because of the nature of our experience, a broad range of stimuli and situations has become associated with those stimuli which are innately set to cause emotions. The reaction to that broad range of stimuli and situations can be filtered by an interposed mindful evaluation. And because of the thoughtful, evaluative filtering process, there is room for variation in the extent and intensity of preset emotional patterns . . . (Damasio, p. 130).

Griffiths is aware that his "natural kinds" threaten to blend together under influence from the phenomena, and he defends the border by two strategies. First, he admits and then minimizes the point I have made. He says,

> It seems clear that emotions [read: affect programs] are sometimes triggered as a result of higher cognitive processes. A complex chain of reasoning may reveal that an entirely novel stimulus is dangerous, and fear ensues (p. 92).

After the brief admission in these two sentences, Griffiths turns away from the point and for several pages discusses affect programs that are "triggered despite, or in opposition to, higher cognitive processes" (p. 92). And note the rather evasive and reducing locution he uses in his admission: "triggered as a result of." The expression is designed to dissociate the higher cognition from the emotion, to suggest that it is somewhat incidental to it – certainly not defined by it – but not even really triggered by it but only triggered as a result of it. Five pages later he comes back to the admission and expands it in such a way as to confirm the distance between the higher cognition and the emotion. Here he rather confusingly says that "In some cases higher cognitive processes may be able to trigger emotional responses directly" (p. 97) and then in a footnote takes this back a bit:

> Creating emotion by imagining emotionally significant stimuli may be an example of the direct effect of central cognitive processes. On the other hand, this may work via the generation of visual and other sensory imagery. It would then be an internal analogue of the triggering of emotion by the visual arts (p. 97, note).

Surely there are cases of autonomically vigorous emotions about objects accessed via higher cognition that are not cases of "imagining emotionally significant stimuli." And what if emotions *are* occasioned by visual and other sensory imagery? Would this suggest, as Griffiths's "on the other hand" seems to imply, that higher cognitive processing is not involved? Many paintings have a narrative background the understanding of which contributes significantly to their emotional impact. Might Griffiths say that such a narrative background, as contrasted with the supposedly lower cognitive mere sensory impression, cannot contribute to the autonomic arousal characteristic of the affect programs? In any case, he goes on:

> The modular system which triggers emotion has an interestingly biased mechanism for learning. The biased nature of the learning mechanism provides a further reason for thinking of *the emotion-triggering system as independent of higher cognitive processes*, since the biases seem to be specific to the learning of emotional responses. The isolation of the triggering system from cognition need not be complete. It may be that rational evaluations of stimuli as emotionally significant can cause the triggering system to be sensitive to those stimuli in future instances. I know of no experimental literature that would allow this to be determined at present. There are also results which seem to show that higher cognitive processes can affect the initiation of emotions in much the same way that they affect the perception of pain (Melzack 1973). . . . Finally,

Ekman's work on cultural display rules suggests that other cognitive processes can block the display of automatic emotional responses by recruiting the bodily systems involved for other purposes (p. 97f, italics added).

So we seem to have four ways that Griffiths admits higher cognition to be possibly involved in the "triggering" of the affect program responses. The first is that affect programs might be triggered via imagination, though it is possible that this is not really higher cognition because such imagining may just produce sensory imagery. Let us clarify his second, third, and fourth possibilities with examples. A man might have an ordinary untutored fear of snakes. Let us say this is unreflective and automatic, thus conforming to the basic paradigm of the affect programs. But then he reads about the poisonous properties of copperheads in his district, thus possibly overlaying his lower cognitive responsiveness with higher cognitive information. Griffiths is admitting that such higher processing of the "stimulus" might enhance the basic responsiveness. The third kind of case would be a higher cognitive assessment, not of the "stimulus" but of the emotion itself. For example, a therapist might tell a client who has a mild phobia of snakes that, because of the therapy he has received, he will respond less violently to a snake the next time he encounters one. Griffiths is admitting that such higher cognitive "input" might actually affect the emotional response. An example of the fourth kind would be a traditional Japanese person whose culture has taught him that it is impolite to look sad. Consequently, when he is impinged upon by a sadness-evoking stimulus, he automatically checks himself after an instantaneous facial expression of sadness and puts on a happy face. Griffiths seems to admit that the effect of the cultural teaching, which is presumably higher cognitive, on the duration of his sad facial expression might thereby reduce other marks (say, blood pressure) of the syndrome.

I submit that the involvement of higher cognitive processes in the panic of our hypothetical professor does not fit any of the four models of involvement that Griffiths admits. He is shy to acknowledge cases in which the very object of an affect program response is wholly and fundamentally presented via higher cognition and could be presented in no other way.[15] To admit this

[15] Besides this kind of case, there seem to be "lower" cognitive emotions that also fail to fit the paradigm. LeDoux says, in a comment that is reminiscent of the affect program thesis, that there is probably no such thing as the neurophysiology of *emotion*, but instead each of the basic emotion types has its own neurochemistry and circuitry. Thus separate investigations must be undertaken for fear, sadness, anger, disgust, and so on. But interestingly, LeDoux mentions a basic emotion that is not included in Ekman's list or in Griffiths's, namely, lust or sexual arousal. An essential mark of the affect programs, as Griffiths defines them following Ekman's work, is that they have a distinctive facial expression that is panculturally recognizable. It seems likely that lust, despite its obvious attractiveness from an evolutionary point of view and the plausibility of supposing it to have a distinctive physiology, gets left off the list of affect programs because it lacks such a facial expression.

kind of case would be to give up the most significant part of the "fracturing" of the concept of emotion that is the fundamental thesis of Griffiths's book.

Perhaps Griffiths will respond that such imagined cases and anecdotal evidence as my panic stricken professor are insufficient to establish the claim that his supposed "natural kind" emotion categories overlap. Damasio, too, offers only an imagined case:

> If you hear of an acquaintance's death, your heart may pound, your mouth dry up, your skin blanch, a section of your gut contract, the muscles in your neck and back tense up while those in your face design a mask of sadness. In either case, there are changes in a number of parameters in the function of viscera (heart, lungs, gut, skin), skeletal muscles (those that are attached to your bones), and endocrine glands (such as the pituitary and adrenals). A number of peptide modulators are released from the brain into the bloodstream. The immune system also is modified rapidly. The baseline activity of smooth muscles in artery walls may increase, and produce contraction and thinning of blood vessels (the result is pallor); or decrease, in which case the smooth muscle would relax and blood vessels dilate (the result is flushing) (Damasio, p. 135).

But the supposition that such cases occur, and occur in abundance in human life, is so much a part of common sense and is so well supported by everyday experience that I should think it is Griffiths who needs to supply empirical evidence that people do *not* experience autonomic arousal in response to objects accessed essentially through higher cognitive processing. He supplies no such empirical backing and indeed seems to try to prevent the reader from looking at the obvious evidence.

Another example of this evasiveness is his discussion of Damasio (Griffiths, pp. 102–106), who, as we have seen, believes that the "lower" parts of the brain, like the amygdala, are susceptible to inputs from the parts of the brain that mediate language and reasoning. Damasio does, admittedly, make a couple of mistakes in his account of the relation between the higher and the lower processes. The first is not quite the mistake of which Griffiths accuses him – that of supposing that all emotions involve visceral activation. Instead, Damasio posits an "as if" circuit that enables the brain to read visceral activation even when there is none, analogous to the pains that amputees sometimes experience in their missing limbs. This allows him to posit that even emotions involving no visceral activation are experienced as "somatic markers" – pleasant or unpleasant somatic sensations. Griffiths's objection that some episodes that fall under the ordinary concept of emotion do not involve any measurable visceral change can be translated into a correct objection, namely that some emotions do not even involve any experience "as if" of visceral change. Another mistake is that Damasio has a somewhat confused concept of *innate* (see Griffiths, pp. 55–64, 104–106). But Damasio's basic thesis that organs like the amygdala can be activated by higher cognitive processes is almost certainly correct

and devastating for the basic thesis of Griffiths's book. One would expect Griffiths to aim all his fire power at *that* thesis, but instead he presents arguments against the weaker spots in Damasio's picture, arguments that create a bit of smoke to keep the reader from seeing clearly that Griffiths did not answer the crucial argument against his view that the affect programs and the higher cognitive emotions are two completely discrete classes of things.

The fact seems to be that some emotions exhibiting the physiological syndromes that Griffiths ascribes to the affect programs are informationally encapsulated, and others are not. Instances of phobic fear are good examples of the former group; our panic-stricken professor is an example of the latter. And there are many intermediate and mixed cases, in which higher and lower cognition work in tandem or the emotion is partially but not wholly susceptible to rational modification. Brain scientists such as Damasio and LeDoux have partial explanations of these phenomena in terms of the modularity of the brain. The phenomena do not lend themselves to Griffiths's neat two-part fracturing of the concept of emotion into discrete natural kinds. For all that we have seen so far, the concept of emotion is more like a mosaic with a variety of "opposites" such as Rorty identified, opposites that do not divide the whole mosaic into two or three separate and complete pictures, but opposites that occur more or less throughout the mosaic. The question remains whether the mosaic has any overall order in it or whether it is just a great, nondescript agglomeration of oppositions of light and dark, of color, and of shape, as Rorty's thesis, on one interpretation, proposes.

I would like to comment on Griffiths's vocabulary before going on to the category of irruptive motivations. Like many biologically oriented psychologists, he is fond of speaking of the objects of emotions as "stimuli" or "triggers." In Griffiths's case, this is perhaps in part a reaction against the propositional attitude theorists, for whom the subject's situation, as it is believed by the subject to be, determines the identity of the emotion (its type-identity, but also its particular identity – see Section 2.5c) and perhaps is constitutive of it. But biologically oriented psychologists who are unconcerned about the menace of conceptual analysis also speak in Griffiths's way because they think of the emotion essentially as behavior rather than as a mental state. Only behavior, after all, will affect the environment, and effects on some piece of the environment (making it ignore you, scaring it off, running from it, planting some of your seed it, etc.) are what is significant for survival and reproduction and thus evolution. But such behavior has to be correlated with environmental variables or occasions – that is to say, stimuli or triggers of the behavior in question. So the biological psychologist is really interested in two things: behavior and behavioral triggers. Now among nonhumans, certain stereotypical situations will pretty reliably trigger the fear response. For example, rats can be reliably predicted to exhibit fear upon perceiving the approach of a cat. But for humans,

the situation is different, as Griffiths notes:

> Emotions in humans . . . are elicited by an enormously wide range of stimuli. A mere light on a control panel may precipitate fear or joy. The only thing which all eliciting situations for fear have in common is the extremely abstract property that, in the light of the organism's past learning history, they can be evaluated as dangerous. This makes it unlikely that each of these situations possesses a range of common features of a fairly immediate perceptual kind (p. 86).

So it is not quite right to call the situation that a human being fears the "trigger." The "trigger," instead, is the situation *as perceived as dangerous*. But the situation as perceived is not the trigger *of the emotion*; it is a subjective state of the subject that "triggers" at best *part* of the emotion – for example, the visceral markers and the behavior. Thus it is really better to think of the perception of the situation in the terms distinctive of the emotion (e.g., dangerous for fear, offensive for anger, disgusting for disgust, etc.) as part of the emotion. It seems to me that this would be a natural way for a brain scientist to think of the matter, if she were thinking just in terms of the brain processes and not in terms of their evolutionary significance: The perception that mediates the onset of the various neural events is itself a neural event and one that may, as we have seen, involve input from very sophisticated, indeed distinctively human, parts of the brain. But as soon as we start thinking of human emotions in such a way that neither the situation that elicits the emotion nor the perception of the situation is the trigger of the emotion, then we see that the same thing is true of the other animals. For them, too, the emotion is not triggered by the situation simply but by the situation as perceived by the animal in a certain way; but the animal's perception of the situation in the distinctively emotional way is not the trigger *of the emotion* but part of it.

d. Higher Cognitive Emotions

According to Griffiths, under the influence of science the vernacular concept of emotion fractures into two parts, the affect programs and the higher cognitive emotions, and this latter category again fractures into two, irruptive motivations and disclaimed actions. In this subsection, first I look at the category of irruptive motivations, and then I look more broadly at higher cognitive emotions. In the next subsection I will address disclaimed actions.

The idea of irruptive motivations comes from Robert Frank's *Passions Within Reason*, which is a development of sociobiology theory. According to Frank, emotions are episodic adaptive departures from means-end reasoning. They are adaptive, despite the fact that they disrupt the ordinary rational pursuit of the subject's goals, because they secure the longer-term success of those goals by their effect on one's conspecifics (namely, other people). The sense of fairness is one of Frank's examples. He has shown

experimentally that people will often refuse to accept an unfair distribution of goods to themselves even when they know the alternative is to receive nothing at all. Out of commitment to being treated fairly, they eschew the (short-term) rational strategy of maximizing their benefits. Similarly, loyalty to a friend who has been fired may lead a person to turn down the job when it is offered to himself. If we look only at the short term, these actions seem irrational. But if a person is known by his fellows to be willing to sacrifice significantly to be treated fairly, he puts his conspecifics on notice that they had better treat him fairly, and this is a long-term gain. Similarly, a person reaps many social advantages by becoming known as a loyal friend. Griffiths points out that such "emotions" as loyalty and a sense of fairness are quite different from the affect programs, in that they are mediated by higher cognition and do not have the same physiological marks. He is attracted to this theory because of its evolutionary bent, though he criticizes it, with other forms of evolutionary psychology, for being insufficiently disciplined in the use of adaptationist explanations. These psychologists speculatively spin "just so" stories that are unrestrained by knowledge of the ecological context in which the trait developed or by interspecies homologies. For example, the sense of justice is explained, as a trait of the species, by the adaptiveness of being fearsome to one's conspecifics when not treated fairly – the conspecifics are deterred from unfairness by anticipatory fear. But if people had the trait of becoming friendly and affectionate when treated unfairly, then the evolutionary psychologist would explain this trait as adaptive, say, by one's conspecifics being wheedled into treating one more fairly by friendliness and affection. Thus it seems that with sufficient imagination, any trait that a species actually *has* can be explained by *some* kind of adaptivity (and there may be competing adaptive "just so" stories that can be told, especially when the ecological context of adaptation is not known). This method of explanation is thus too "easy," not sufficiently controlled by the data. By contrast, the affect programs can be explained evolutionarily, not just as adaptive but as having homologues in related species whose ecological contexts can be known. Thus, by contrast with the scientific work that has been done on the affect programs, Griffiths goes so far as to say that we do not know what the higher cognitive emotions are (see pp.229, 241) because we do not know what their underlying homeostatic causal mechanism is.

We may wonder whether loyalty and a sense of fairness even belong in the vernacular concept of an emotion. But some of the things that Frank attempts to explain using the strategy I have just described do seem to belong in the category: Jealousy, guilt, shame, and vengefulness are examples. Griffiths, however, is offering the category of irruptive motivations as one of three that exhaust the field covered by the ordinary concept of emotion (the third being disclaimed actions). So we must ask whether all higher cognitive emotions that are not ways of faking emotion so as to justify some otherwise unjustifiable action are irruptive motivations.

Griffiths glosses "irruptive motivational states" as "states that interfere with the smooth unfolding of plans designed to secure our long-term goals" (p. 246). But there seem to be plenty of higher cognitive emotions that do not interfere with the unfolding of our plans but are a direct consequence of our plans and fit smoothly into them. Consider the gratitude a person feels when, endeavoring to accomplish some task, a friend comes along and offers welcome help. This is certainly a higher cognitive emotion, may have very little of the physiological markers characteristic of the affect programs, and is not a fake (let us say). The accomplishing of the task is enhanced, rather than impeded, by the feelings and expression of gratitude. Or consider hope. Someone in the family has been injured, and several phone calls have been made without securing medical help. Then finally, one call gets through to the doctor, and a feeling of hope comes over the family. This emotion depends on a sophisticated conceptual understanding of the situation; it is not fake; but neither is it disruptive of anybody's long-term goals. I submit that countless instances of higher cognitive emotions are neither irruptive motivations nor social pretenses.

We might also wonder whether nonfake higher cognitive emotions are all motivational. Take the quiet satisfaction a person feels in having solved a difficult and important engineering problem. It is clearly an emotion, yet it does not seem to lead to any action (other than further contemplation of the accomplishment). Furthermore, the kind of examples in which Frank specializes seem tendentious. Goal disruption is relative to some particular goal. But we typically have many goals, of different kinds, between which we may be torn by the situations of life. Loyalty to one's sacked friend may "disrupt" one's goal of getting the best job one can, but it may *express* the goal of friendship. Why suppose that friendship is not a perfectly real goal of people who are loyal to their friends? And to suppose that such loyalty is irrational is to accept a very narrow conception of rationality as narrow self-interest. Guilt may "disrupt" the "goal" of feeling good about oneself, but it may express the perfectly rational "goal" of being a morally upright person.

Some higher cognitive emotions that are not fake are disruptive and motivational, but not all are, by a long shot. I submit that the concept of an irruptive motivation is far too narrow to do the categorial work that Griffiths assigns it. Besides this, there is the conceptual difficulty that the category seems to reflect a very particular ideology of rationality: It is only by this contestable standard that these emotions can be construed as short-term irrational or disruptive.

e. Disclaimed Actions

The third category into which Griffiths thinks the vernacular concept of emotion fractures under scientific pressure consists of "emotions" that are unconscious strategic pretenses imitative of emotions belonging in the other

two categories but lacking the "passivity" characteristic of the affect programs and the irruptive motivations, pretenses that are reinforced by local culture and serve the purpose of licensing behaviors that would not be socially acceptable if performed without the excuse provided by the fake "passivity" of the fake emotion. Griffiths is not fully confident that this third category should be allowed, and nowhere does he claim that disclaimed actions are a natural kind. But in claiming that this category is discrete in that its states are strategic and therefore not passive, Griffiths does invite us to propose counterexamples and critical questions about the discreteness. I shall argue that the inference "strategic therefore not passive" is fallacious, that countless human emotions that seem to fall in the categories of affect programs and higher cognitive emotions have an intentional element, and that being intentional does not mark a neat division among kinds of emotions, but on the contrary many human emotions fall on a *continuum* of intentionalness. We can also question whether intentional emotions are all strategic, whether the subject is always unconscious of his own role in getting them up, and whether they are always "reinforced by local culture."

The passiveness that Griffiths thinks is characteristic of the affect programs and the irruptive motivations must not be merely that they are not intentionally brought on by their subject. If only that were meant, then got-up emotions would be uninterestingly unpassive by definition. If passivity is not to be thus trivialized, it must be an attribute of the subject under the influence of the emotion, and it must consist in something like inertia, in the emotion's not being purely under the volition of the subject and instead having a "life of its own," causing the subject to feel and behave in certain ways. In other words, for a subject to be passive to an emotion is for the subject to be *in its grip*. Part of this may be affect program type grip – heart beating faster, and so on; part of it may be motivational – you feel like hitting somebody, and so on; part of it may be what Stocker calls affectivity[16] – feeling a certain way characteristic of the emotion, where this feeling is not to be identified with bodily sensation. Passivity means that the subject has a sense of *being* compelled in such ways *by* the emotion. Thus the question whether everything that is done intentionally lacks the passivity characteristic of emotions is an empirical question.

Non-emotional analogues of this are not contrary to deliberation and strategic intention. For example, when you sled down a hill, you may carefully choose a position from which to start the descent, strategizing how to get the effect you want (a slow smooth safe ride, a fast exciting bumpy one, etc.) and pushing off hard or gently. But once on your way you may be carried along, without volitional control, by the sled. Or you may be partially in control (say, you can steer it a bit) and partially not in control (you can't

[16] See Michael Stocker and Elizabeth Hegeman, *Valuing Emotions* (Cambridge, England: Cambridge University Press, 1996), pp. 17–55 and *passim*.

stop it). Human experience suggests that many got-up emotions (perhaps all of them, to some degree) are like this: They take hold of you, some very gently and some violently.

Dostoevsky's underground man is especially prone to getting up emotions and is unusually astute about his own states and the relationship between his deliberate production of the emotion and its "passivity." His remarks suggest that his got-up emotions are quite capable of taking him over in the way characteristic of real emotions.

> I used to invent my own adventures, I used to devise my own life for myself, so as to be able to carry on somehow. How many times, for instance, used I to take offence without rhyme or reason, deliberately; and of course I realised very well that I had taken offence at nothing, that the whole thing was just a piece of play-acting, but in the end I would work myself up into such a state that I would be offended in good earnest. All my life I felt drawn to play such tricks, so that in the end I simply lost control of myself. Another time I tried hard to fall in love. This happened to me twice, as a matter of fact. And I can assure you, gentlemen, I suffered terribly. In my heart of hearts, of course, I did not believe that I was suffering, I'd even sneer at myself in a vague sort of way, but I suffered agonies none the less, suffered in the most genuine manner imaginable, as though I were really in love. I was jealous. I made scenes. And all because I was so confoundedly bored, gentlemen, all because I was so horribly bored.[17]

The underground man's got-up emotions are strategic, with at least two aims: making life miserable for others and dispelling his own boredom. But the strategy works best when the emotion is not *mere* play-acting, but the passivity of the emotion supervenes, as it often does in his case. His emotions are a clear counterexample to Griffiths's "strategic, therefore no passivity" inference. He is also a counterexample to the principle that disclaimed action emotions require that the subject be unconscious of the strategizing. The underground man is often transparent in the getting-up of his emotions up and in why he is doing it. We might also wonder to what extent the underground man's strategies are sanctioned or reinforced by local culture. Dostoevsky himself thinks that the underground man is a cultural product,[18] but it is not clear that the culture functions as a sanctioner or reinforcer of the sort that Griffiths envisions.

The method of acting developed by Constantin Stanislavski[19] presupposes that people can be trained to bring themselves into the states of emotion of the characters they are playing on stage. For Stanislavski this means that they do not merely reproduce the facial and bodily movements and

[17] *Notes from the Underground*, in *The Best Short Stories of Dostoevsky*, translated with an Introduction by David Magarshack (New York: Modern Library, no date), p. 122.

[18] See the author's note at the beginning of the story.

[19] *An Actor Prepares*, translated E. R. Hapgood (New York: Routledge, 1948). I thank Sarah Borden and Lance Wilcox for alerting me to Stanislavski's writings.

vocal inflections that a character in a given emotional state would produce, but the actors themselves *feel* the emotions. Stanislavski believes that without this "inner experience" (p. 164) the acted behavioral expression will always be somewhat wooden and artificial: "All external production is formal, cold, and pointless if it is not motivated from within" (*ibid.*). If we apply this insight about acting to the kind of cases of disclaimed action that Griffiths cites, then, far from "strategic" implying "not passive," "not passive" (that is, not genuinely felt as an inner motivation, a lived experience of emotion) implies "not strategic" (that is, not very likely to be a successful strategy). If romantic love is a social role that licenses certain behaviors, then, on Stanislavski's principle, the license will not be very effective if "there is no state of love to explain love behavior" (Griffiths, p. 246). Stanislavski thinks that human beings have memories of a variety of emotional experiences that they can learn to trigger and apply to new situations. The memories are somewhat schematic, so that in triggering them one is not just reliving the original experience, but is having a new experience that fits the schema. As an actor practices a part, he may deliberately recall emotional situations of his own past that are formally similar to the situations of his character in the play. The physical movements of the actor on stage, if they are well-executed and attended by receptivity to memory, can trigger the appropriate emotional memories and thus reproduce the experience of the character. Another stimulus is "the text of the play, the implications of thought and feeling that underlie it and affect the inter-relationship of the actors . . .," and another is "all the external stimuli that surround us on the stage, in the form of settings, arrangement of furniture, lighting, sound and other effects, which are calculated to create an illusion of real life and its living *moods*" (p. 191, Stanislavski's italics).

Paul Ekman, Robert Levenson, and Wallace Friesen have supplied some empirical evidence that supports Stanislavski's theory. They instructed actors to contract their facial muscles in ways that (unknown to the actors) are characteristic of expression of various emotions, and they monitored their heart rate and finger temperatures. They found that autonomic responses characteristic of the expressed emotions occurred with no other provocation than such "acting."[20] But regardless of the accuracy of Stanislavski's theory about *how* professionally enacted emotions are produced, he is clearly right that actors do sometimes experience their characters' emotions, thus reproducing the "passivity" involved in them, and that they do this deliberately.

A case in which an emotion comes on partly spontaneously and partly intentionally is found in Jane Austen's *Sense and Sensibility*. The subject is Marianne Dashwood ("sensibility"), a partisan of the romantic ideology of strong feelings. Willoughby, with whom Marianne has recently fallen madly

[20] "Autonomic Nervous System Activity Distinguishes Among Emotions." *Science* **221** (1983): 1208–1210.

in love, has suddenly and inexplicably departed from the county for an indefinite period. She is wrenched with grief. Her sister Elinor ("sense") tries in vain to understand the departure and the relationship:

> But whatever might be the particulars of their separation, her sister's affliction was indubitable; and she thought with the tenderest compassion of that violent sorrow which Marianne was in all probability *not merely giving way to as a relief, but feeding and encouraging as a duty* (Chapter 15, italics added).

It is implausible to suggest that Marianne's affliction is nothing but a pretense, even if it is supposed to be an unconscious and therefore sincere one. Her violent sorrow shows the passivity characteristic of genuine emotion; and we may suppose that her facial expression is panculturally recognizable, and physiological measurements would suggest that she is in the affect program state that Ekman calls "sadness" – this despite the fact that romantic love does not seem to be pancultural. Furthermore, her emotion is clearly a higher cognitive one because her ability to conceptualize *an indefinitely long period of time* is essential to its situational object. If we accept Elinor's description of Marianne as intensifying her romantic sadness out of a sense of duty, then we can see that the intention with which a subject brings on his emotion is not always strategic (unless we think of people's dutiful actions as strategies for getting their duty done).

In mixing spontaneity and intention, Marianne's emotion seems to be typical of an enormous number of the emotions of sophisticated human beings. In subtle ways we engineer our emotions (or deengineer them), sometimes strategically and sometimes under ideological (religious, antireligious, political, ethical, aesthetic) inspiration (or pressure). All such emotions seem to violate the supposedly neat fracture lines that Griffiths thinks science produces in the concept of emotion. Griffiths will no doubt respond that my data are all anecdotal or, worse, literary. But the examples are true to our experience. If someone proposes that experience is fundamentally inaccurate, then we need strong empirical justification for the claim. To justify the particular fracture lines that Griffiths posits, we need empirical studies showing that no emotion that requires higher cognition for its onset has the physiological characteristics of the so-called affect programs. We need studies showing that all emotions that are neither fake nor informationally encapsulated are both disruptive of the subject's long-range goals and motivational. We need studies showing that no emotions that the subject intentionally brings on himself have the character of "passivity." Griffiths cites no study of any of these kinds.

The impression I have tried to convey in this and the preceding two subsections is the complexity of the phenomena that we call emotions, a criss-crossing of properties in the field that foils any effort to divide the emotions into neat, mutually exclusive nonoverlapping categories. It is a complexity to which Amélie Rorty drew attention and which she tried to

reduce in a very different way. If present-day biology seems to demand such a division (because of its relative success in dealing with certain instances of the affect programs and its awkwardness in dealing with much of anything else), it seems to me that this may indicate that biology has its explanatory limits. It is good at some kinds of things and not so good at others, and it is not very good at explaining a great deal that belongs to the field of emotion.