Chapter 1: introduction: the affect program theory of emotions

There probably is no scientifically appropriate class of things referred to by our term "emotion." Such disparate phenomena — fear, guilt, shame, melancholy, and so on — are grouped under this term that it is dubious that they share anything but a family resemblance. But particular emotions are another matter altogether. There is good reason to believe that different sciences can make quite compelling sense of a more fine-grained differentiation of affects. My task in this book is to reveal some of the important and neglected lessons of some of the emotions for the philosophy and sciences of mind, and this task can be accomplished with just a working characterization of a few of these. More importantly, there is a compelling theory of some emotions which has far-reaching implications for the philosophy and sciences of mind. This is the affect program theory. Using a version of this theory as a guide both to what phenomena we will be concerned with, and to the nature of these phenomena, will allow us to avoid fundamental confusions and to provide richer results.

The affect program theory is the view that some emotions are pancultural syndromes enabled by inherited biological capabilities. By calling them "syndromes," we mean to point out that they are coordinated collections of complex biological responses which occur together. These emotions will be characterized by several features, including at least physiological responses, such as autonomic body responses, and stereotypical associated behaviors, such as facial expressions but also relational behaviors. I will call the emotions that are taken to fall under the affect program theory "basic emotions," just so that we have some way to refer to them.i This is a very general

formulation of the affect program theory; however, with some small elaboration in this chapter, it will be sufficient to allow me to draw some very important lessons about the nature of mind.

This theory is meant to describe only some of the things that we might call an "emotion." In part as a result of this, there is plenty of room for controversy regarding whether this is a proper theory of emotion. For example, some theorists have argued that conscious experience is a necessary element of an emotion (Clore 1994), whereas this is not the case on the affect program theory. Thus, one might argue that the affect program theory does not properly describe the emotions as the normal speaker means to refer to them. The affect program theory is an empirical theory; it is not beholden to fit exactly our folk use of affect terms, or our folk theory about affects (see Griffiths 1997). Ultimately, the defense of the affect program theory must rest on how well (1) it usefully defines and distinguishes the various affects, and (2) it explains and predicts the relevant phenomena. Defending the theory's utility to explain and predict the relevant phenomena is done throughout this book, by way of applying the theory and showing how it can offer powerful new ways to think about some of the problems of mind. Defending whether the theory offers a useful way to categorize the affects is something I will do in this chapter and the next. My approach will be to examine some of the features of affects that other scholars have singled out as necessary or sufficient or perhaps even just important to emotions and other affects. We will see that our best scientific understanding of these features reveals that they are either consistent with the affect program theory, or they are themselves not appropriate ways to ground a theory of affect. This will also allow me to review the scientific evidence and theoretical reasons that lay the foundations for a view

of mind that is quite different from most of those that characterize contemporary philosophy of mind.

Although scientists have tended to be more careful, and usually provide sufficiently operational notions of the emotions and other affective states they study, until recently (until, e.g., Griffiths 1997) there has been scandalously little concern among philosophers (even philosophers of emotion) for clarifying their taxonomic presuppositions. This oversight is not innocent, since it fosters both an extremely error-prone armchair theorizing, sometimes even armchair neuropsychology, and also vagueness and confusions which can result in question-begging and pernicious ambiguities.

Most philosophy of emotion has proceeded in one of three ways. In recent years it has been most common for emotions to be investigated through the use of emotion terms. This is an approach which is sometimes taken to an extreme by those who endorse the position that the conceptual analysis of ordinary language is all that is needed to understand emotions, or by the social constructionists, who see culture — of which they take language to be the most important and revealing element — as the creator of emotions. Paul Griffiths has effectively criticized the former approach to emotion studies, pointing out that in this context it has been based upon philosophical presuppositions which are now largely debunked (1997: 21ff). I will criticize a strong social constructionist approach in chapter 4. A second method for philosophizing about emotions, more revealed in the lack of an explicit method, is to take emotions as primitives open to reliable introspection; not surprisingly, this approach usually yields the view that emotions are fundamentally cognitive. But taking emotions as having

properties which are somehow obvious inevitably leads to begging of all the important questions; emotions are introspected to have just the qualities needed to support whatever theory is at hand. I shall review some cases that show how psychologists and neural scientists have discovered some very surprising things about our everyday emotions; things which would certainly fail to be noticed by introspection. Introspection also results in subjective characterizations that are hard or impossible to pin down. Without some, even if rough, prior and objective (that is, third-person, open to observation) characterization of the things we are discussing, much of this work on emotions can be useless. A third approach is to simply define emotions and work with these definitions; this also has traditionally yielded cognitive approaches. Defining emotions up front in some cognitive form would be, of course, quite acceptable, if this were not usually followed by sweeping generalizations that reach beyond the scope of the class of phenomena picked out by the definition. As it stands, all too often we find a theorist start with a definition of emotions that is strongly cognitive, and then make claims about all emotions, surreptitiously slipping in the assumption that all of what others call "emotions" fall under the definition of emotions as cognitive. We therefore either need to be extremely careful not to erroneously generalize from our definition, or we need to characterize (at least some) emotions in some sense which is in part guided by empirical data and also allows us to formulate the core questions about emotions. I will take the latter route, beginning with a broad characterization of affects that is not by-definition cognitive, and then exploring how we can build our way to a characterization of some emotions which will let us learn some lessons from them.

<1> A general notion of affect

It will be useful to start with a more general characterization of affect. This will give us a chance to place the relevant emotions in relation to things like pleasure or mood. There is little agreement upon terminology for emotions and other affects in philosophy, psychology, or any other of the cognitive sciences. In general, terms like "emotion" and "affect" are used synonymously. However, for most of us (at least in the English-speaking world), paradigm emotions include fear, anger, joy, sadness, and disgust. At the same time, some people consider moods to be emotions, including thus long term states that have very different motivational features than, say, terror. And philosophers will talk about the importance of emotions to rationality, seemingly grouping desire and other more general conative states together under the term "emotion." Given that such a disparate group of things can be labeled "emotions," we need to draw some distinctions between these phenomena. Here I shall try to avoid confusions by using "affect" as a general term, and desires, emotions, moods and other states will classify as types of affects.

I still need to characterize affect in some positive way. The working definition I propose is: affects are body states that are motivational. (Throughout this book, I will take body states to include neural states; when I want to draw attention to the body independent of the central nervous system, I will use the term "extended body.") This is not in itself very enlightening, since motivation is not a little mysterious. But the principal feature of these motivations is that they are internal physical states of an organism that cause it to perform an action if the organism is not inhibited by different

motivations or otherwise constrained. The relation of inhibition by other motivations, and also the notion of constraint, although both intuitively clear, are very hard to specify. Without a better account of what it is to inhibit or constrain a motivation, this characterization might be too vague if we meant to explore the nature of affect per se. But the claim that the affects are types of body states is sufficient to distinguish this notion of affect from many of the competing notions; in particular, it commits us to a realist theory of motivations (in contrast to, for example, ascriptivist notions of desire, such as I discuss below and in chapter 3). Furthermore, this is a claim for type-identity: the body states that motivate are instances of a recognizable type. Since it will be sufficient to have a working notion of just a certain class of emotions, I will take motivation as a primitive; however, this notion, as it is involved with the basic emotions that will be my concern here, will be developed at more length in the coming chapters. In the meantime, this definition makes it clear that I link affects to actions.

There are a few other features I need to clarify in this characterization of affects.

<2> Affect is characterized in a functional way

Affects include desires, pleasures, emotions, and moods. We should note that these things are quite distinct in the physiological and, in particular, neural structures that underlie their function; we should not expect to find a single brain system for all motivation. Furthermore, when they are cognitive, affects can include significant input from not only subcortical brain areas but also from cortical polymodal and supramodal areas — more simply put: a lot of the brain, including areas seemingly dedicated to more

abstract thought, can (but need not) become involved in the affect. Thus, as occurs with many biological functions, we should expect some of the brain and body substrates of affects to be distributed. All of these distinctions reveal that this notion of affects is a functional characterization that may not in any simple way reduce to a physical one.ii We may indeed find that the neural underpinnings, for example, of some particular affects can be quite clearly mapped out; but the concept of affects in general is unlikely to have such a common characterization.

Two other things should be noted about this characterization of affects. First, although I believe that they are necessarily motivational, pains are often understood in neuroscience as somatosensory phenomena that activate a motivational system. We could use "pain" in a broader sense to be understood to include the activation of the motivational systems that neuroscientists take the somatosensory aspects of pain it to activate; but, given that nothing here depends on it, I will instead avoid expending effort on what could be contentious issue. I will not require that pains be counted as affects. Second, moods pose special difficulties; since moods will only be a passing concern here, I will not try to characterize them at more length. As a working notion, we can think of moods as long term affective states, perhaps even long term emotions; as such, their motivational aspect is revealed more as a long-term and consistent alteration in motivation (relative to the subject when not in that mood).

<2> Affects are not all bivalent/monodimensional states

Many have suggested that affects are states that are either negative or positive appraisals (of something, such as the organism's situation). Thus, it is extremely common in psychology to group emotions into groups with "negative" and "positive" valence. Similarly, some philosophers have defined emotions as belief states coupled with some bivalent feature or one-dimensional magnitude meant to capture the affective aspect of the emotion; Greenspan (1988) uses comfort/discomfort as this feature, while many others (e.g., Marks 1982) assume desire is this feature. I will not respect these uses of the term "affect" because I believe that they are ultimately unhelpful; and, at the very least, although they may be valuable when used to describe some affects, such uses are not useful as broad characterizations of all affects. For example, the notion of an appraisal or state being "positive" is too vague. What makes an appraisal positive? Ultimately, if the notion of a positive or negative appraisal is not to be vacuous, it must either yield some measurable feature of the body, or, better yet, it must reveal something about the kind of behavior that such an appraisal results in (such as approach or avoidance). One supposes that joy, for example, is positive (as per colloquial usage of "positive") and that it leads to approach (in some sense). But what about anger and fear? Colloquial usage would make them negative; but one can lead to approach of the emotion's object (in attack), the other to retreat from it (in flight). Given such distinct behaviors, the categories just do not explain anything. Similarly for comfort and discomfort. Suppose anger and fear are uncomfortable. What does this tell us about the behaviors that would result? That we seek to avoid them? But it seems, at least prima facie, that we sometimes seek these emotions, through art (revenge films include bad guys who are there specifically to raise our ire, and frightening movies garner audiences

because they are frightening) or activities (like seeking fights or riding a roller coaster). Or does it mean that once we have the emotion we seek to get out of it? But, again, if a movie-goer or a mountain climber is even partly motivated by the thrill of fear, their behavior is inconsistent with such a supposition (they stay in the theater, or they keep climbing). Pleasure/displeasure, comfort/discomfort, and positive/negative, and various degrees of satisfaction of a desire, are all too crude to tell us anything interesting about many of the emotions and the behaviors that typify them.

Note that I am not arguing here against the use, by neuroscientists and others, of activation and inhibition (and cognate notions) of behaviors as general explanatory posits (e.g., Gray 1991); I am rejecting the use of (usually far more general) one-dimensional measures for taxonomizing emotions and other affects into, say, the positive group or the negative group. Another way of making the same point is to note that such monodimensional categorizing threatens to be far too impoverished for explaining data. It can result in such a reductive simplification that effects of the phenomena involved can be lost as they are pressed onto a single measure.iii One solution to this kind of simplification is to introduce a host of bivalent appraisals for each emotion; this is a strategy taken by Ortony et al (1988) in their discussion of the cognitive origins or causes of emotions. They argue that emotions are bivalent reactions concerned with three aspects of the world: events, agents, or objects (18). But, of course, multiplying the number of dimensions in a model can distinguish any number of states; so before we accept a complex of bivalent appraisals or of some monodimensional features, we need some independent reason to accept the dimensions that are being offered. Here, we shall see that dropping the very notion of bivalent appraisals, and related notions, loses us

nothing. The term "affect" will be used in a way that does not presuppose some bivalent measures such as discussed above.

<2> Affects are occurrent states, not dispositions

Affect terms can all be used in a dispositional sense. If we say that Tony desires chocolate, or that Eric is angry at his landlord, we could mean at least two things in each case. We could mean that the person in question is in a particular body state, or we could mean that he tends to be in that body state, given the right conditions. The former I will call an occurrent affect, and the latter a disposition to affect. iv Thus, in ordinary discourse a sentence like "Eric is an angry person" can be ambiguous; it could mean that Eric is right now angry, or it can mean that Eric is the kind of person who is often angry. Similarly, one might say that Eric has been angry at his landlord for years, but of course it is not the case that anyone can be in an occurrent state of anger for that long a period of time. Instead, we mean that when reminded of his landlord or confronted with his landlord, Eric usually becomes angry. We might also mean that the beliefs and values that Eric holds that cause him to be angry at his landlord — say, the belief that his landlord is charging him too much money, and the high value he places on being treated justly, and so on — are still held by Eric, which should have as a consequence that when he attends to these things he has an occurrent state of anger as a result. Or, Tony can be said to have a disposition to desire chocolate if he desires chocolate often, or if he desires chocolate whenever he sees it. But Tony only has an occurrent desire for chocolate if he is actually in a state of desiring chocolate. Disposition to emotions and other affects are

of particular importance to our normal discourse because we use them in attributions of temperamentv and other affective personality traits: a sybaritic person may be someone who has a disposition to desire to ingest chocolates and to pursue the experience of various other pleasures, a choleric person is someone who has a disposition to be angry. However, the concept of disposition to affects is (at least as I am using the term here) derived from the concept of occurrent affect, and does not admit of many of the features that occurrent emotions have (for example, there is no sense in arguing whether a disposition to affect is a propositional attitude or not — this could at best mean that the occurrent affect for which one has a disposition is itself a propositional attitude). I shall hereafter mean an occurrent affect by any affect term.

<2> Affects are real physical states, not ascribed explanations

There is a related notion of affect which can be held by someone who denies that there are occurrent affects, and holds that talk about affects and about disposition to affects are both just a convenient gloss for dispositions to behavior. On such a view, attributions of affects may not correspond to an actual body state but rather might just be a kind of logical construction relating actions and beliefs.vi Thus, Adam might ascend the steps to his front door in a single leap out of habit. It may be that there is no significant sense in which Adam has a kind of body state which corresponds to the desire to leap up to the door; rather, he may just do it out of habit, without any need to chose between this option and the option of taking the steps one at a time. However, one might still say that Adam "desires" to leap the three steps in a single bound, and simply mean

by this that he believes (if he were queried) that he can get to the door that way and furthermore he does in fact get to the door that way. We then might attribute the "desire" as a kind of relation between the relevant belief or beliefs and the relevant action. One who is very skeptical about affects being actual body states in any significant sense might advocate that all or many such affects are just kinds of logical attributions. There are in fact measurable occurrent states that seem to correspond to instances of desire-like states (though it is dubious that there is any generic motivational state like the philosopher's notion of desire), but I need not defend this claim here, since my goal is to develop a theory of some of the emotions — emotions for which it is uncontroversial that there are strongly related physiological and brain states. We need only note, then, that affect terms as they are used here will not be meant as mere logical relations between belief and action or between any other mental states or actions; what they stand for must necessarily include actual (in principle measurable) body states which are best identified as states resulting from or constituting the affect.

<1> Distinguishing features of affects

Ultimately, I believe that we shall do best to fix a theory of basic emotions and other affects on a developed scientific understanding of the neural circuits that enable those affects. Thus, our best criteria to identify affects will include such factors as Panksepp (1998) utilizes: he holds that "The most compelling evidence for the existence of such systems is our ability to evoke discrete emotional behaviors and states using localized electrical and chemical stimulation of the brain" (52). In fact, such evidence

often reveals quite definite neural structures, some of which offer very compelling neuroanatomical evidence in favor of the affect program theory. I shall refer to some of the relevant neuroscientific evidence throughout my discussions of the basic emotions and other affects. However, since my task here is in part to relate the affect program theory to common-sense notions of emotion, including to the kinds of features that have traditionally come in for much conceptual analysis and therefore which have been of concern to philosophers, I will begin with a number of directly observable or introspectable features; these features are also a good starting place because some of them are likely essentially linked to the functional role of the relevant affects.

Such possible distinguishing features of occurrent affects that have interested scientists and philosophers are their physiological state, conscious experience, associated actions, and relations to cognitive content.[vii] We might also add to this list the relative temporal duration of the affect: generally, it seems that affects that are not moods or emotions do not last as long as emotions, and that emotions last less long than moods. One might hold that two affects are indistinguishable on their physiology, but could be distinguished according to duration (sadness and depression, for example, might be such a case). There is a significant body of literature on stress that is concerned, in fact, with duration of some affects. For my purposes here, however, this research will not be taken to be sufficient to characterize the emotions.[viii] Here I will remain agnostic about all the possible meanings of differences in duration. Instead, I will turn next to the first three of these four features. Since in the next chapter I will be discussing the cognitivist theories of emotion (the view that emotions are in some part constituted by, or at least require,

beliefs or other propositional attitudes), I will leave a discussion about cognitive content for that chapter.

<2> Physiological state

Affects, especially some emotions, have noticeable, and measurable, physiological correlates. For example, a large body of research reveals that some forms of decision making (and thus, presumably, very basic forms of affects) result in the, often very subtle, autonomic changes measured by electrodermal recordings of skin conductance (e.g., Damasio 1994). For emotions, many more measurable physiological changes occur. Depending upon the intensity of the emotion, these can include changes in autonomic functions, such as heart rate, blood pressure, respiration, sweating, trembling, and other features; hormonal changes; changes in body temperature; and of course changes in neural function as measurable by EEG (Frijda 1986: 124-175).

For a long time, it has been controversial to suppose that some of these changes were distinct for particular emotions. It has often been seen as an important element of a cognitive theory of emotion to hold that the physiological changes accompanying an emotion amount to a kind of undifferentiated excitation, and that cognitive contents were needed to distinguish anger from fear, happiness from sadness, and so on (a source often cited in support of this view is Schachter and Singer 1962ix). However, much of the previous failure to clearly establish distinguishing physiological profiles for emotions or other affects appear now to largely have arisen because of the inadequacy of past measuring techniques. Although the claim remains controversial, evidence is growing

for the view that autonomic activity distinguishes among at least some emotions. Ekman, Levenson, and Friesen (1983) have found, for example, that discrimination between several emotions (fear, anger, sadness, disgust, surprise, and happiness) was possible just by observing temperature and heart rate changes from baseline measurements (that is, measurements of the subject when presumably not experiencing the emotion). Since these are measurements from a baseline, this study does not establish that we can actually identify one of these emotions in a subject on first observation, but it does at least show that we can distinguish the emotion from some others when several measurements are available. These experiments were done with actors, but later found to work with normal subjects (Levenson et al 1990). They also worked not only for directed facial action (asking subjects to form the expression of an emotion) but for reliving (that is, recalling, thinking through) an emotional experience; and results from many other researchers is consistent with these findings (see Levenson 1992 and 1994 for a review). More research is needed in this area as there are some outstanding questions that remain,x and the experimental difficulties are great (generating fear, sadness, joy and so on in laboratory conditions is not easy), but these results are substantial and encouraging: they show that a significant number of the emotions may be distinguishable from each other by these autonomic features alone.

These results do not yet allow us to identify emotions by their physiological effects or constituents. But these kinds of investigations at least provide compelling evidence that there are reliable physiological changes that accompany some affects; for the emotions that we will be concerned with here, there is sufficient evidence that these affects necessarily include physiological responses such as changes in temperature, heart

rate, and other features — even when the subject is having a relatively weak emotional experience, and even when the subject may be unaware of any such changes. Many cognitivists will deny that emotions necessarily have these correlates. In such a case, we can just be disagreeing about the semantics of our terms: these cognitivists take emotions to be mental contents, perhaps social relations, and these other features are incidental. But, as I will show in the next chapter, such a position is not only inconsistent with the scientific evidence, it leaves us unable to distinguish emotions from other kinds of mental states. The claim that measurable physiological changes are necessary — leaving open whether they are sufficient to identify the relevant emotions — is important because such changes are sufficient to distinguish emotions from some other states with which some like to conflate emotion, such as belief. Furthermore, the autonomic patterns and related physiological changes are surely part of the phenomenal experience of some emotions. And, these physiological responses are probably also essentially connected to relational actions and other affective behaviors. At the very least, we must explain or take into consideration these physiological features if we are to have a satisfactory theory of emotions.

If the physiological changes accompanying an emotion are necessary, but perhaps not sufficient, to identify that emotion, we must turn next to the three features of conscious experience, associated actions, and relations to content in order to get a more complete understanding.

<2> Conscious experience of affects

Affects like anger, fear, despair, pleasure, and many others can have distinct conscious experiences. It might then seem that affects all are necessarily accompanied by a conscious experience; and, in fact, many scientists and philosophers assert that emotions must be conscious. There is much ambiguity in the term "conscious" here, one that has recently come under much analysis by philosophers (I will return to this in chapter 9). However, in this section I am concerned with the notions of consciousness that scientists tend to use; intuitively, a process is conscious if the subject is aware of it, in some sense reflecting upon it, and can use that awareness in directing or performing some action. In deference to contemporary uses, we will call this sense of "consciousness" working consciousness whenever there is a threat of ambiguity.xi What it is to be aware of a state is not clear, and there certainly are mental states of which the subject is not aware but which influence working-conscious action. This lack of clarity alone casts grave doubts upon the idea that we can gain any definitive understanding of emotions by asserting that they are "conscious" or by otherwise finding a role for consciousness in them. Thus, in order to try to ground my discussion of consciousness and emotions, I will have to find some criteria for something being conscious. One sign of working consciousness is that the agent can, barring any deficiencies (such as brain damage making speech impossible, etc.), report on the state. This is a too-strong criterion, and it does not get to what the notion of working consciousness seems to be aiming for (that is, I grant that the ability to report on a state is not the same as being aware of it). However, it is, at least, relatively clear. Furthermore, it comes close to capturing, I believe, what is really motivating many who insist that emotions must be working conscious: a notion that emotions play a part in our rational and deliberative

control of our activities. So, for the sake of clarity, I shall use in this section the very strong criterion that a process is working conscious if a subject can report upon that process (I am leaving vague what counts as a report; this should be acceptable since the examples discussed below are clearly instances of inability to give different kinds of report).

If we are to retain the idea that motivation is the quintessential feature of affects, then not all affects are working conscious. Instead, we have strong evidence that there are affective states which are effectively motivating but of which the agent is not able to report — not even indirectly. One of the most interesting examples is found in the mere exposure effect, as primarily championed by Zajonc. Much research has established that people tend to prefer familiar stimuli, even when they fail to properly recognize that stimuli (see Zajonc 1968, 1980). What Zajonc and his colleague found was that subjects could form preferences for certain stimuli to which they were subjected for extremely short durations (e.g., tens of milliseconds), making it extremely unlikely that they have performed the kind of complex cognitive processing necessary for categorizing and memorizing the stimuli for later recognition of a kind sufficient for a declarative reporrt (Kunst-Wilson and Zajonc 1980). By first showing subjects pictures of shapes, and then later forcing them to choose among a number of shapes including but not limited to those to which they had been exposed, results showed that subjects would choose amongst the shapes at chance when forced to pick out the shapes they had seen before, but would show a significant preference for stimuli that they saw before when asked to chose the shapes they preferred. These results can also be shown for some nonhuman animals, such as mammals (Hill 1978).

An everyday example with similar import might be the use of polygraph machines, the so-called "lie detectors." Lie detectors measure skin conductance response, which changes as a result of activity in sweat glands and which appears to be well-correlated with other physiological changes. As we have noted, physiological body activity, including autonomic activity, is one of the distinctive features of at least some affects. What is interesting is that normal subjects show measurable galvanic skin conductance changes to certain kinds of plausibly affective situations — such as to a situation in which they want to deceive and be undetected and are, perhaps as a result of their awareness of their potential loss from being detected, experiencing some affective reaction of which they themselves need not be conscious nor over which they have any conscious control.

But one might argue that the affective states seen in mere exposure and in subtle skin conductance differences are not emotions, and that although some affects can fail to be conscious, emotions are always conscious. However, it seems possible that emotions are in fact capable of being unconscious. First, an emotion might be unconscious in the sense that one has an impaired ability to understand or describe the emotion. Such seems to occur in some cases of alexythemia (Sifneos 1972). These subjects show an impairment in both the verbal and non-verbal recognition of emotions (Lane et al 1996), and this impairment can in severe cases extend to their own emotions. Sometimes these subjects report that they are experiencing some kind of an emotion, show some of the stereotypical behavior of an emotion (e.g., weeping), but are unable to say what caused the behavior (Nemiah and Sifneos 1970) or to properly categorize it. There is also some evidence that some alexithymics can have abnormally large autonomic responses to

emotion-generating stimuli (Martin and Pihl 1985, Papciak et al 1985). The best explanations of alexithymia are of the form that an emotion is occurring, but that the individual is not properly aware of it (perhaps in a way analogous to blindsight; see Lane et al 1997) or is unable to properly categorize it (perhaps because of a failure to have developed a cognitive skill to recognize emotions; see Lane and Schwartz 1987).xii If by an emotion being conscious we mean that the subject can recognize and properly categorize the occurrence of an emotion in herself, then the alexythemic subject has an unconscious emotion.

---------- EXPLANATION BOX 1.1 ABOUT HERE ----------

Second, there is evidence supporting the hypothesis that some phobias arise because some individuals have an inherited predisposition to fear certain stimuli (including concrete objects), and that this predisposition allows for fear reactions that are unconscious (see Seligman 1971). Thus, results similar to the mere exposure effect have been found for fear by Öhman, Dimberg, and Esteves (1989), Öhman (1988), and Öhman and Soares (1993, 1994). In these experiments, subjects have demonstrated skin conductance responses for fear conditioned stimuli that are presented so quickly, and with masking, that they are not consciously recognized. For example, in Öhman and Soares (1993), subjects were tested with fear-relevant images (snakes and spiders), along with neutral images (flowers or mushrooms), some of each of which were used in fear conditioning; following the conditioning, these stimuli were shown for short durations and followed with a mask (a neutral stimulus that follows the initial stimulus, and which

therefore interferes with any conscious memory of the initial stimulus); skin conductance responses were then shown to be strong only for the conditioned fearful stimuli. This strongly suggests that unconscious processing of some kind is sufficient to cause fear responses. These results were also shown to be independent of where in the visual field the stimulus was presented, which is consistent with the process being subcortical since no lateralization (as occurs with many cognitive, cortical processes) of the ability is observed. Similar results were found using images of angry or happy faces, utilizing aversive conditioning only for the former (Esteves, Dimberg, and Öhman 1994) (attention can have effects on these results, but the subject are not conscious of the stimuli in that they cannot identify, even in forced-choice tests, the stimuli after exposure). These kind results provide strong evidence for at least fear conditioning and for fearful or phobic responses occurring unconsciously.

Third, there is a significant body of psychoanalytic literature dedicated to the idea that emotions can be unconscious, and that they can still play an important role in shaping actions by, and in the psychoanalytic explanation of the behavior of, the subject. It is unclear, however, to what degree and in what senses these emotions are unconscious. Are they dispositions to emotions, which lead to occasional occurrent emotions of which the subject is unaware? Are they, as Freud apparently held, not emotions but emotionally relevant unconscious beliefs? Or is it that the subject sometimes has occurrent emotions and it is the cause of these of which the subject is unaware? It will not be my place to try to answer these questions here, but only to note that some of these kinds of explanations require that unconscious emotions be possible. If any proves to be a powerful explanatory tool, then it is reason to posit unconscious emotions.

Fourth, for some theorists, the underlying notion of an emotion being working conscious seems to be that we somehow know why we are having the emotion, that we are aware not only of the emotion but also of its cause or at least its object (and, on some views, its cause should be its object). We can refine our criterion in such a case to include that the subject can report not only that they are having an emotion, but also why they are having it; or at least that when having an emotion the subject is aware of the object and cause of it. If this is required for one's notion of what it is for an emotion to be conscious, then the view that emotions can be unconscious has some valuable supporting evidence to be found in neuropsychology. Working with split-brain patients in the 1970s, Gazzaniga and LeDoux were able to show a very clear sense in which emotions were not, in this sense, conscious. These split brain patients are people who have had a commissurotomy, a surgical procedure to cut the commissure, a bundle of nerves that connects the two neocortical hemispheres of the brain. This procedure is used as a last resort treatment for some forms of epilepsy. But neuropsychologists have long known that the two hemispheres of the brain have specialized functions. What happens if you separate one of the primary links between them? Gazzaniga was able to show that subtle deficits can be revealed under controlled conditions. A stimulus can be shown to one side of the brain, leaving the other side of the brain, in some senses, unaware. For example, the language centers of most people are in the left hemisphere. Showing a figure just to the right side of the brain (done by placing it only in the left hand side of the field of vision) can result in the object being (in some senses, defined in the respective experiments) recognized, but the split-brain subject being unable to say what the thing is. LeDoux and Gazzaniga used this same approach to study affects. They could show the

right side of the brain an affective stimulus. The specialization of language being (usually) in the left hemisphere of the neocortex, the subject was wholly unable to verbally report on what the stimulus was. But the affective import of the stimulus seemed to somehow be "leaking" to the left neocortex. The subject, wholly unconscious of what the stimulus had been (in the sense of being unable to report on it), could properly categorize it under some value terms as "bad" and "good." This at least shows that affective characterizations or related value judgments can be made in a way that is unavailable for report. Here is a sample case:

[when] a word was [shown only] to the right hemisphere and P.S. [one of the subjects] was instructed to perform the action described by the word, his reaction to the word kiss proved revealing. Although the left hemisphere of this adolescent boy did not see the word, immediately after kiss was exposed to the mute right hemisphere, the left blurted out, "Hey, no way, no way. You've got to be kidding." When asked what it was that he was not going to do, he was unable to tell us. Later, we presented kiss to the left hemisphere and a similar response occurred: "No way. I'm not going to kiss you guys." However, this time the speaking half-brain knew what the word was. In both instances, the command kiss elicited an emotional reaction that was detected by the verbal system of the left hemisphere, and the overt verbal response of the left hemisphere was basically the same, regardless of whether the command was presented to the right or left half-brain. (Gazzaniga and LeDoux 1978: 151)

The researchers conclude that this result "is inconsistent with the currently accepted cognitive theory of emotion" because in P.S. "the left hemisphere appeared to have experienced a directionally specific emotion in the absence of a cognition" (152). That is, the affective reaction was significantly directed — it resulted in or was a withdrawal from a suggested course of action — and the subject is clearly aware of something. However, the subject is not aware of the affect in a way that enables him to identify it's cause or object; it would appear, at least, that the kind of conscious awareness that a cognitive theory of emotion requires was not present. Using my terminology, the behavior here is not necessarily revealing a basic emotion: it may require only what we are calling "affect." But it does at least show that some strong affective reactions, plausibly related to emotions, are unconscious in this sense. The next case is more relevant to emotions.

Gazzaniga also found spontaneous emotional reactions of laughter unavailable to report. In the following passage, "the machine" is the apparatus used to ensure that visual stimuli are seen only in one side of the visual field, and thus only get to the opposite hemisphere:

When a pin-up was flashed without warning to the right hemisphere of [the subject], amongst a series of more routine stimuli, she first said, upon being asked by the examiner, that she saw nothing, but then broke into a hearty grin and chuckle. When queried as to what was funny, she said that she didn't know, that the "machine was funny, or something." When the picture was flashed at the left hemisphere she laughed

too, and quickly reported the picture as being a nude woman. Using a different modality (olfaction), Gordon and Sperry (1968) recently confirmed this kind of result.

Neither hemisphere in [another subject] found the nude overtly funny (he was 51 at the time of testing), but did find other testing situations humorous. In one test of tactile learning capacity, using the left hand, [this subject] broke out laughing when feeling one member of the stimulus pair. The particular stimulus consisted of a tack nailed into the middle of a wooden square block. Every time he felt it, he would pick it up and twirl the block about the axis and would chuckle heartily when doing so. When asked what was funny he would say, "I don't know, something in my left hand I guess." (Gazzaniga 1970: 105-106)

If laughter is properly an expression of an emotion, then that emotion is occurrent but unavailable to the relevant kind of introspection for these patients. Note that I do not endorse, and these observations do not require us to conclude, that emotions are cortical (that is, that the neural centers that underlie their function are in the neocortex) and lateralized (that is, that the underlying neocortical center is specific to one side of the brain); rather, for my concerns here the point is that the kinds of capabilities that constitute working consciousness in this stronger sense, or at least that offer criteria for its existence, are cortical and lateralized phenomena. These split-brain studies show failures in working consciousness that differentiate some of its features from emotions or other affects.

Defining affects in terms of their conscious role is therefore unlikely to be a strategy that succeeds well in identifying affects, or otherwise helping us to understand

them, since some of them can be unconscious and still have behavioral effects other than conscious reports. Nonetheless, given that some affects are sometimes not working conscious, it still remains that certain affects seem distinguishable from other affects by the nature of their phenomenal experience when there is such a working-conscious awareness of the experience. This is particularly true of the common emotions: rage, joy, sadness, fear, and shame — to pick just a few examples — seem to have feelings (when they are conscious) specific to the emotion (or, at least, specific enough to distinguish the emotion from other kinds of states, like belief), so that we may find it convenient to use their conscious experience as one of their distinguishing features. Should we then define some affects, such as emotions, in terms of their phenomenal experience? This strategy has several stumbling blocks. There is the problem, already observed, that some affects can be unconscious. But, supposedly the suppressed affect has effects on the subject, and these are the kind of effects one would expect of that affect. If we are able to identify unconscious occurrent anger with a working-conscious instance of occurrent anger, or any suppressed affect with its working-conscious counterpart, then the common element must be something other than the phenomenal experience of the affect, since those properties per se are just freely spinning wheels without the working conscious awareness of them. If there are unconscious emotions we thus cannot have as a defining feature of affects their phenomenal features alone. Another problem is that, like the "feeling theory of affect" which has long been in disrepute, treating emotions as conscious states characterized by one's awareness of the experience does not explain much. Ultimately, philosophers and scientists want to understand how emotions relate to behavior, and what role they play in our mental lives,

and specifying how they "feel" does little to further this goal. A related problem is that reference to phenomenal experience does not give us any objective (that is, third person) criteria with which to distinguish these emotions. But we certainly do properly recognize emotions in others, and if we are to study affects in a scientific (that is, third person) way we will need some objective criteria with which to work.

The conscious experience of an affect, although important, cannot be a fundamental feature used to define emotions.

<2> Associated actions

Another approach to understanding and categorizing emotions is to look at the kinds of behaviors that they cause, or with which they are associated. This is not to say that emotion concepts are nothing but useful ways of grouping together disparate classes of behaviors; on the contrary, looking at emotional behavior has also provided evidence that many of them are highly associated with stereotypical, pancultural behaviors; and this in turn should be viewed as evidence that the behaviors themselves are caused by biologically-based, inheritable capabilities.

Some of the most compelling evidence for some pancultural human emotions has come from studies of facial expression. It was Darwin who first argued at length that facial expressions of emotions are evolved emotion-expressing behaviors. In more recent times strong evidence has been gathered that Darwin was correct. Eibl-Eibesfeldt studied the facial expressions of children born deaf and blind, some with extensive brain damage. He discovered that these children showed spontaneous signs of emotions such as smiling

when playing or sitting in the sun, laughing when playing, and crying when in an unfamiliar environment (1973; see also Fulcher 1942). Some of these children had severe cognitive deficits, and none were able to see or hear the emotional expressions of others, so it is highly implausible that they learned these behaviors.

Cross-cultural studies of facial expressions have found evidence of high cross-cultural correlations. These kinds of studies were made in a thorough manner by Ekman and colleagues (Ekman, Sorenson, and Friesen 1969), who sought to get as pure a cross-cultural study as was possible; they created a set of thirty photographs of facial expressions that they felt expressed six emotions that other psychologists had proposed as basic (happiness, surprise, fear, anger, disgust, and sadness). They then showed the photographs to college students in the US, Brazil, and Japan, and volunteers in New Guinea and Borneo. The six emotion terms were translated between languages and then the subjects were asked to group the pictures under the terms. A very significant degree of agreement was usually found — higher for some emotions than others, and for some cultures than for others, but in general there is an unmistakably significant degree of agreement. Ekman and Friesen recreated this experiment (1971) working with the Fore of New Guinea, a cultural group relatively isolated from the rest of the world, and found agreements again ranging from 64 percent (for fear) to 92 percent (for happiness). This work, and related work (Izard 1971), supports the view that human facial expression of some emotions is pancultural.

These results have some interesting supporting evidence in neuroanatomy. There are two distinct neural pathways that control facial movements. One is through the pyramidal tract, and the other through the phylogenetically older extra-pyramidal tract. It

seems that emotional facial expressions are controlled by the older, extra-pyramidal tract. This is evident when damage to the motor cortex that impairs motor control of the face (as often occurs in hemiparalysis) can (when the damage is localized to the motor control area) sometimes be spontaneously overcome when the unfortunate subject expresses emotion. In other words, a stroke victim might be unable to smile on the paralyzed side of the face when so commanded, but might smile involuntarily and normally at a joke. Conversely, damage to the extra-pyramidal tract could leave voluntary control intact but result in the loss of all spontaneous emotional facial expression (Rinn 1984).

These findings suggest that emotional facial expression is pancultural because of inheritable, evolved neural structures that are shared by all, or at least many, human beings. There is also interesting corroboratory evidence available for this view in studies of nonhuman primates. Research by Miller, Caul, and Mirsky has shown that the facial expressions of rhesus monkeys can transmit significant information to other monkeys, and though monkeys raised in isolation fail to recognize the meaning of the facial cues of other monkeys as well as do the monkeys raised in a social setting, these isolated monkeys still showed facial affective cues that other monkeys could recognize and properly understand (1971). This research goes some way to showing that our near evolutionary cousins share with us the having of innate facial expressions of affect, and that the innate expressions are therefore highly likely to have evolved in a common ancestor.xiii

These results all find surprising support in some of the studies by Ekman of facial expressions among Japanese and American college students (1980). In the experiments, each student was left alone to watch films, some of which were stressful, and some of

which were not. Their facial expressions were recorded, and these recordings measured by researchers who did not know what films the subjects were watching. When the students were alone, both Japanese and American students showed significantly similar facial expressions. In some cases, however, a lab-coated individual was put into the room with the subjects. In these cases, as expected, Japanese students altered their expressions much more, smiling more and showing less stress. This is consistent with the facial expressions being pancultural, but wholly amenable to different display rules. More interestingly, for those cases where researchers were in the room with the subjects: "Examining these videotapes in slow motion it was possible to observe sometimes the actual sequencing in which one movement (a smile for example) would be superimposed over another muscle action (such as a nose wrinkle, or lower lip depressor)" (1980: 94). In other words, the evidence suggests that the pancultural facial expression is being generated but then promptly suppressed. Note that this is also very suggestive of a two-track view of these emotional expressions: a potentially subcognitive emotion causes the facial expression, perhaps primarily through the extra-pyramidal tract, and a slower, secondary, cognitive appraisal suppresses it.

This is also consistent with the use of surface electromyographic recordings (EMGs) in studies of emotion (see Tassinary and Cacioppo 1992). EMGs measure muscle action potentials in, for example, the face — that is, neural stimulation of facial muscles. They can detect these muscle action potentials even if they fail to result in any change in facial expression, for example if they are too weak or too short in duration to cause a muscular action. EMGs provide a tool for psychophysiologists to measure facial reactions to emotion-eliciting stimuli even when no observable facial expression change

occurs. The underlying method is guided by the belief that emotions can cause muscle action potentials which are not under conscious control or awareness of the subject.

Emotional behaviors are more than just facial expressions, however. Perhaps one of the most compelling accounts of the use of emotion concepts is found in Hebb's classic 1946 article on the recognition of emotion. Hebb reviews an experiment at a primate laboratory where for two years the scientists working with the primates were not allowed to use emotion terms to describe the animals' behaviors. Instead, they had to keep records which described only what the animals did at one time or another. What Hebb discovered is that describing different chimpanzees without using emotion terms left people unable to really convey the sense of the character of the different primates. One could not tell, just from looking over the records of past events — described painstakingly without "anthropomorphic terms" — what the animal was like or would behave like:

All that resulted was an almost endless series of specific acts in which no order or meaning could be found. On the other hand, by the use of frankly anthropomorphic concepts of emotion and attitude one could quickly and easily describe the peculiarities of the individual animals, and with this information a newcomer to the staff could handle the animals as he could not safely otherwise. Whatever the anthropomorphic terminology may seem to imply about conscious states in the chimpanzee, it provides an intelligible and practical guide to behavior. The objective categorization therefore missed something in the behavior of the chimpanzees that the ill-defined categories of emotion

and the like did not — some order, or relationship between the isolated acts that is essential to comprehension of the behavior. (1946: 88)

A pragmatist should be satisfied on this observation alone that emotions are genuine scientific entities. Someone of a more realist bent might rightly argue that Hebb's conclusion is true because some emotions lead to, or are in some way linked to, actions which are specific to and explicable by these emotions.

Hebb's observations should remind us of the strategy of the ethologist. The ethologist looks to find the patterns of behavior in animals. If there are patterns that occur again and again, and if these patterns can be found in isolated groups and even in closely related but different species, then this is some evidence for a homologous behavior. The ethologist is not therefore much distinct from the evolutionary biologist, utilizing the concept of homology for behaviors as well as for anatomical structures (where homologous behaviors would presumably arise from, and ultimately be explained by reference to, homologous structures). The ethologist's method applies to humans as well (see Eibl-Eibesfeldt 1989). Evidence that some emotion expressions are pancultural, that the structures allowing for the expression are inheritable, and that certain patterns of reoccurring behavior are inexplicable (not regularly predictable) without emotion concepts all point toward the primary thesis that some emotions can be identified via their homologous associated actions.

This makes sense of the presence of emotions in other nonhuman animals. Our primary means of recognizing fear in a rat, anger in a dog, surprise in a cat, and so on, is through the behaviors that they show in such states. Scientists regularly use these criteria

(and others, such as autonomic responses) to study emotions in nonhuman animals. It is difficult to see how else we are going to understand these claims except through the identification of shared kinds of behavior.

Some cognitivists about emotion have argued that observations of behavior fail to provide any evidence for emotions in nonhuman animals, and therefore fail to support theories like the affect program theory. Ortony et al have claimed that:

... it is tempting to suppose that animals experience fear. However, such attributions are typically based on observations of behaviors (aggressive behavior or avoidance behavior), which turn out to be dissociated from the emotional states to which they are presumed to be linked.... It would be a relatively straightforward matter to program a robot to exhibit aggressive or avoidance behavior toward certain objects or classes of objects, yet, if having done so one were to claim that one had produced the emotions of anger or fear in the machine, one would be scoffed at by the scientific community, and rightly so. (1988: 27-28)

There are at least two errors in this argument. First, it is not the case that we posit that there are emotions in nonhuman animals just because we observe simple behaviors. An ethologist who sees a bird leaving a branch to fly to another is not about to claim it fled the oak in fear in order to attack a maple out of anger. We posit that there are emotions in nonhuman animals because it is the best explanation for a very large body of evidence. This evidence includes, but is not limited to, the existence of behavioral patterns which are in particular ways both flexible and inflexible, so that they reveal the pursuit of a kind

of action (see chapter 3); which are observed again and again; which can be best described as fulfilling the functions that we suppose in our theories that these emotions fulfill, or even that we ascertain the emotions in ourselves fulfill; which can reliably be described as being caused by eliciting conditions consistent with that function; which are reliably accompanied by expressive behaviors; which include autonomic and other physiological changes which are special to the emotion, and perhaps even shared by us; and which (most importantly!) in some animals are caused or constituted by neural structures which have homologs in the human neuroanatomy of emotion. Thus, behavior is a crucial element, but it does not stand alone. Second, it is a patent falsehood that it is a "straightforward matter to program a robot to exhibit aggressive or avoidance behavior toward certain objects or classes of objects." It is a major accomplishment to get a robot to navigate a small, unchanging, and extremely simplified environment. To get a robot to actually recognize, effectively track, and pursue a resistant (say, a moving or even fleeing) object so that the robot could effectively attack it is truly the kind of engineering beyond, or at least at the very limit of, contemporary AI and robotics. Of course, it is a straightforward matter to program a robot that on a flat surface in a featureless environment moves towards or away from a light, for example (the kind of "behaviors" sometimes referred to as "Braitenberg behaviors"; see Braitenberg 1984). But this cannot be the kind of thing that is meant by "aggressive or avoidance behavior," because the very thing at issue here is the attribution to nonhuman animals of emotions had by humans; and so no respectable argument for the presence of these emotions in nonhuman animals would depend on counting such simplistic "behaviors" as examples of aggression or avoidance (and, as noted, the attack or avoidance behaviors of most animals is

extremely sophisticated). Furthermore, this kind of reasoning may well be backwards; to program a robot that can exhibit effective behaviors like aggression and avoidance with the kind of flexibility that even a relatively simpler animal (such as an insect) reveals — to actually engineer an autonomous agent — may best be accomplished by working with a robust theory of affects, and then attempting to engineer an affective agent (I shall indulge in some speculations in this direction in chapter 12). Finally, to suppose that it is a simple matter to program these behaviors may be an instance of a common fallacy — what I will call the cognitive autonomy fallacy — that what is not cognitive is simple and inflexible, while what is cognitive is complex and flexible and the wellspring of autonomy. I will return to this point several times.

The view that some emotions can be identified through the actions with which they are associated is perhaps merely a consequence of my definition: since affects are motivations, then the principal method we have for discerning and distinguishing them is through the behavior they motivate. We can always keep in mind, however, a realist (as a philosopher would call it) criterion: when we identify an affect, we are identifying a genuine physical state of an individual organism, and if it later turns out that there is no such significant (that is, measurable) state, or the behavior was best explained in some other way, then we were wrong to so identify the state. In the cases of things like preferences, the motivation is very general (let us assume, for a moment, that there is a state corresponding to "preference"). If a subject S prefers to do some action A, then we are saying little more than that S is in a motivational state which has as an effect that they will A, ceteris paribus (when it is possible, when they are not constrained, and when there is no stronger motivation to do something inconsistent with A). But other affective

states are much more structured. We can understand fear by supposing that if subject S fears some object O, then S will flee from O — with the same ceteris paribus clause. Some emotions, it seems, are characterized specifically by the complex behavior that they have as a consequence — what psychologists sometimes call "relational actions," since they are explicitly concerned with relations to other things (Frijda 1986: 14-24).

<1> The affect program theory

Some of the things that we call "emotions" appear to be a collection of things: physiological responses, stereotypical actions, and perhaps even normal cognitive roles. Instead of reductively explaining these emotions in terms of one of these features, I will adopt the naturalistic theory that tries to respect all of them: the affect program theory. This theory is not favored by philosophers, nor psychologists who work on the social-end of their discipline, but in various forms it is quite common to psychobiologists, neuropsychologists, and others who concern themselves with the biology of emotion. I adapt the notion from Ekman, who took the term from Silvan Tomkins:

For there to be such complexity and organization in various response systems, there must be some central direction. The term affect program refers to a mechanism that stores the patterns for these complex organized responses, and which when set off directs their occurrence.... The organization of response systems dictated by the affect program has a genetic basis but is influenced also by experience. The skeletal, facial, vocal, autonomic,

and central nervous system changes that occur initially and quickly for one or another emotion, we presume to be in largest part given, not acquired. (Ekman 1980: 82)

By "affect program," Ekman means to refer to only some aspects of the emotions in question. He argues that an emotion is made of an affect program along with a response system, an appraiser, and elicitors (1980: 86-87). In a sense, this is of course correct, and a weak form of cognitivism about emotions is tantamount to the view that all of these things are normally present in emotions but they need not all be. I will therefore here just use the term "affect program theory" to refer to the whole syndrome, recognizing that the cognitive elements are in humans quite common, but unnecessary, and that the physiological and behavior consequences are themselves necessary.

The idea of emotions as affect programs best explained by reference to our evolutionary heritage is perhaps most indebted to the research of Paul MacLean (e.g., 1990). MacLean introduced the "triune brain" hypothesis, in which the brain is seen as having three systems, hierarchically arranged, each of which is to some degree independent of the others and which corresponds to a definite stage of evolutionary development. These systems are the "reptilian brain," the paleomammalian or limbic brain, and the neomammalian neocortex. On this model, many affects are reptilian or limbic system adaptive programs, that in humans can operate to varying degrees independently of our neocortical systems.

The neuroscientist Jaak Panksepp also offers a compelling approach to the basic emotions which is consistent with the affect program theory. He has offered six criteria that distinguish the basic emotional systems:

1. The underlying circuits are genetically predetermined and designed to respond unconditionally to stimuli arising from major life-challenging circumstances.
2. These circuits organize diverse behaviors by activating or inhibiting motor subroutines and concurrent autonomic-hormonal changes that have proved adaptive in the face of such life-challenging circumstances during the evolutionary history of the species.
3. Emotive circuits change the sensitivities of sensory systems that are relevant for the behavioral sequences that have been aroused.
4. Neural activity of emotive systems outlasts the precipitating circumstances.
5. Emotive circuits can come under the conditional control of emotionally neutral environmental stimuli.
6. Emotive circuits have reciprocal interactions with the brain mechanisms that elaborate higher decision-making processes and consciousness. (1998: 49)

What these various approaches share is a common recognition that some emotions are complex, coordinated events that include motor programs or subroutines, that evolved and are recognizable in homologous form in related organisms, and that are fundamentally enabled in neural circuits. For my purposes here, one of the most fruitful features of the basic emotions, as understood in the affect program theory, is the action or motor programs that in part constitute some of them.

<2> The central role of action and the parsimony of the affect program theory

The linking of emotions to actions is widely accepted. Nico H. Frijda claims that "Emotions are changes in readiness for action as such... or changes in cognitive readiness.... or changes in readiness for modifying or establishing relationships with the environment... or changes in readiness for specific concern-satisfying activities" (1986: 466). More strongly, he has written: "It will be clear that 'action tendency' and 'emotion' are one and the same thing" (71). The psychobiologist Robert Plutchik has argued that "an emotion is a patterned bodily reaction of either protection, destruction, reproduction, deprivation, incorporation, rejection, exploration or orientation, or some combination of these, which is brought about by a stimulus" (1970: 12). More recently, he has added that "emotions are complex chains of events with stabilizing loops that tend to produce some kind of behavioral homeostasis.... [The] physiological changes [that accompany an emotion] have the character of anticipatory reactions associated with various types of exertions or impulses, such as the urge to explore, to attack, to retreat, or to mate" (1994: 100). So that

From an evolutionary point of view one can conceptualize emotions as certain types of adaptive behaviors that can be identified in lower [sic] animals as well as in human. These adaptive patterns have evolved to deal with basic survival issues in all organisms, such as dealing with prey and predator, potential mate and stranger, nourishing objects and toxins. Such patterns involve approach or avoidance reactions, fight and flight reactions, attachment and loss reactions, and riddance or ejection reactions. (229)

Silvan Tomkins claims that emotions are "innately patterned responses" and these "affect programs" are stored in subcortical brain centers (1970: 108). Richard Lazarus argues that emotions result from primary appraisal of a situation, and a secondary appraisal results in a coping action. And, as noted above, Panksepp advocates a psychobiological theory of some emotions in which they arise from neural circuits and facilitate adaptive behaviors; these neural circuits "are genetically hard-wired and designed to respond unconditionally to stimuli arising from major life-challenging circumstances" and they "organize behavior by activating or inhibiting classes of related actions (and concurrent autonomic/hormonal changes) that have proved adaptive in the face of those types of life-challenging circumstances during the evolutionary history of the species" (1982: 411). Howard Leventhal has presented a perceptual motor theory of emotions, in which "There is a basic set of stimulus-sensitive expressive-motor templates, each of which generates a different emotional experience and expressive-motor behavior" (1984: 127). I advocate, and will assume here, the hypothesis that basic emotions have as an essential element a motor program.

What is the motor program that is part of the affect program of some emotions? This is an empirical question, but here I can clarify the notion, draw some likely conclusions about its evolution, and warn off likely misunderstandings of the term "program." The program need only be functionally specified for my purposes, but it surely is (primarily) instantiated in a neural system. Once activated, this action program will, if not actively inhibited, result in the emotional behavior. Strictly speaking, the functional definition of the action program therefore has the action as a consequence — like a functional definition of motor cortex activity, for example, can have motor activity

as a consequence.xiv Thus, on this view, given an occurrent basic emotion, it is not the emotional action but the common lack of it, or the modification of it, that requires additional theoretical posits. This is all consistent with the compelling working hypothesis that some emotions evolved from innate behavioral responses — that is, what ultimately amounts to motor programs — in ancestors of the emoting agent. The term "program" is perhaps unfortunate, but I use it because I know of no clear alternative. The motor program is not meant to be a simplistically deterministic list of discrete symbolic instructions, such as a computer program written in Java, for example. It is rather a dynamic capability. A rat running from a fearful stimulus might take a different path each time it flees — but it still may always consistently flee. Many brain systems are perhaps best thought of as dynamical systems (see Port and van Gelder 1995), and like many dynamical systems result in output that is most conveniently described in terms of a range of possible continuous trajectories moving through a state space — which, compared to a computer program, has the flavor of a kind of qualitative, as opposed to quantitative, description.

With this general notion of motor programs in place, the affect program theory yields a bonus of increased parsimony in our theorizing. As we saw, many theories of emotion (including some cognitive theories) share the supposition that it is an essential feature of emotions that they have some kind of significant relation to action; the most widespread agreement is that the emotions are at least some kind of disposition or tendency. Although "disposition" takes on the sound of a substantial and well-placed primitive concept in much action theory, a disposition is a mysterious entity, and provides not a proper part of a theory but rather a debt to be discharged. Present

understanding of the human mind and brain are not sufficient to expect a successful theory of all our disposition talk, and so much or most of our disposition concepts and related concepts are merely placeholders for the possibility of the relevant action. However, I have suggested an inversion of the usual explanations: we should take the emotional action as primary, and either the failure to act, or the cognitive guidance of action, as secondary. Since we do have general theories of how inhibitions can work,xv and since cognition is already an existing problem, there is some theoretical gain in this approach. Every debt we can pay off is, after all, a net gain in our theoretical finances.

<2> Evolution, innateness, and inheritability

The affect program theory will ultimately be verified and fully developed as the relevant neural systems are identified and understood. However, from a functional and from a psychoevolutionary perspective, the most distinguishing feature of an affect program is the behavior that, at least in part, constitutes it. Presumably, like the facial expressions that accompany and express some basic emotions, the more complex relational action patterns that characterize some basic emotions started as motor programs that evolved into inheritable patterns of behavior. As some of the species having these motor programs evolved ("towards" us, for example), some of these behaviors remained, although they became subject to alteration and inhibition via new capabilities that accrued to the species involved. In ourselves, these action programs can be occurrent — one might say, "running" — but result in diverse or even no overt behavior. Thus, the program that makes up an occurrent basic emotion, I claim, is in part the occurrence of

the relevant behaviors (in the broad neuroscientific sense) themselves; and for at least some of the basic emotions, this includes some relational action. The relational action of a basic emotion is a consequence of the occurrent action program if the action program is not inhibited. Similarly, most other features of an affect program can also best be explained by reference to their role in the behavior of the emotion.

But I have been rough with the evolutionary claim about the affect program. This is partly because the conclusions I aim to draw in this book are largely independent of the variations that I gloss over. Thus, how "universal" the relevant affects are is a concern I hope to pass over in the interest of avoiding a set of important but distinct philosophical problems. For my purposes here, any significant portion of the relevant populations having some of these features is going to be sufficient. Thus, I will hold only that the basic emotions are biologically based capabilities (that is, the structures which allow them to occur can be described by a biological science — above all, neuroscience), that they are pancultural (that they arise in every culture, even if not in every individual), and that they are inheritable (the reason they occur in individuals in every culture is because some people inherit this capability). Maintaining only these presuppositions should allow me to avoid such issues as, for example, the degree to which the inheritability of the basic emotions is "innate" or a result of the inheritance of common environments. It is fair to say that no feature of our neuroanatomy is not shaped by learning, and I certainly would deny a claim that basic emotions come prepackaged at birth. But whether affect programs are so very determined by inherited characteristics that they would occur in a recognizable form in radically different environments, or whether instead a significant degree of their inheritability arises because certain environmental features are pancultural

and these help determine the program, is (though very interesting) not relevant to the discussions that follow. Similarly, whether the elements of the emotion syndromes are generated and coordinated by a central neural program, or whether they just occur together because of reliable environmental conditions (and thus, for example, could be controlled by several neural systems that could potentially operate individually, were certain unusual environmental conditions to occur), need not be answered here. I do believe that the affect programs arise from centralized neural programs, but otherwise the issue is one I leave to future empirical research. (For a discussion of these issues see Griffiths 1997.)

<2> Which emotions are basic?

I will call all and only the emotions that are pancultural and that fall under the affect program theory the basic emotions. But there remains disagreement about what these emotions are. Ekman and others involved in facial studies have included fear, anger, joy, sadness, and disgust (Ekman and Friesen 1971). Panksepp doubts that disgust is a basic emotion; he believes that the basic emotions include at least SEEKING, FEAR, RAGE, and PANIC (a social distress system), LUST, CARE, and PLAY (Panksepp's use of capitalization is meant to draw attention to the fact that these are technical terms, related to, but still potentially distinct from, our usual uses of these terms; see Panksepp 1998). Some have also found preliminary evidence that there are pancultural expressions of contempt (Ekman 1988), and embarrassment and shame (Keltner 1995). But since fear and anger are in the intersection of all such lists (such as also Izard 1971, Plutchik

1980; see Kemper 1987 for a review of such attempted lists), for my purposes in this book I shall ensure that each argument regarding the import of basic emotions can be made with this subset alone. We can otherwise remain agnostic about the exact set of basic emotions. For the record, I opine that the union of both Ekman's and Panksepp's lists identify (not necessarily all) basic emotions: fear, anger, joy, sadness, disgust, seeking/curiosity, social distress, lust, care, and play.

<2> Some hypotheses concerning function and eliciting conditions

In arguing that some basic emotions are in part constituted by action programs, I have endorsed a view that these basic emotions have specifiable functions. That is, for example, if part of fear is the action program of flight, then flight is a function of fear; and if part of anger is the action program to attack, then attack is a function of anger. Although it will not be necessary for many of the arguments that follow in this book, it will at times be useful to refer to both potential roles and also eliciting conditions of the basic emotions. These are separate issues, strictly speaking; and yet, one should expect that functions that are type-specific to a basic emotion have eliciting conditions that are also type-specific. Thus, if a function of fear is to motivate a flight from a dangerous object, then we expect that a dangerous object would be a typical eliciting condition.

There is growing evidence that there are some universal eliciting conditions for basic emotions, and for other affects (Boucher and Brandt 1981; Scherer and Walbott 1986; Scherer et al 1986). The general patterns revealed in these and other studies are quite familiar. Ekman and Friesen (1975) identify: an actual or a threat of harm as an

elicitor for fear; loss of an object for sadness; something repulsive for disgust; and frustration, a physical threat, insult, a violation of one's values, or someone's anger directed at oneself being causes of anger. Lazarus (1991) offers a taxonomy of "Core Relational Themes" for various emotions; these help define both function and eliciting conditions. They include: a demeaning offense against me and mine for anger; facing an immediate, concrete, and overwhelming physical danger for fear; having experienced an irrevocable loss for sadness; taking in or being too close to an indigestible object or idea (metaphorically speaking) for disgust; making reasonable progress toward the realization of a goal for happiness; and many others (122).

These and other accounts suggest that, for some of the basic emotions, an abstract characterization of function and eliciting conditions is possible that will be consistent with many of the contemporary theories. Keeping with the idea that I will be able to work just with fear and anger as typical emotions, I will suggest the following:

Fear functions to motivate flight from a threat, and is elicited by the perception of a threat.

Anger functions to motivate an attack against a defeasible enemy, and is elicited by the perception that a defeasible enemy has harmed or intends to harm the organism or something the organism values.

This list is obviously short; I could attempt an account of the functions and elicitors for many other affects (e.g., disgust functions to motivate the expelling of, or withdrawal

from, potential toxins, and is elicited by the perception that something is both potentially digestible and is a toxin). But the actual function and universal eliciting conditions of basic emotions and other affects is an empirical matter, and will require additional empirical investigation. This partial list will be sufficient to make a few points regarding function and eliciting conditions in later chapters, and so I will end with the hypothesis that these two accounts are correct.xvi

Chapter 2: the case against cognitivism

In chapter 1, I observed that one potential feature distinguishing affects is their cognitive contents (or perhaps their relations to these contents). This would be an approach that is consistent with the various cognitivist theories of emotion. My purpose in this chapter is to address cognitivism, and show that it is an untenable view of the basic emotions if it is meant to define them, or otherwise explain their necessary nature. In recent years, many criticisms of cognitivism about emotions have been made (see Deigh 1994, de Sousa 1987, Gordon 1987, Griffiths 1989 and 1997, Stocker 1987, Stocker and Hegeman 1996). These various criticisms have not, however, touched upon the important scientific evidence, especially from neural science, that is inconsistent with cognitivism about emotions. In this chapter, I utilize a sampling of this evidence to explore why cognitivism is inadequate. This will also, as in chapter 1, provide an opportunity to support my overarching themes: the affect program theory, a hierarchical and bottom-up view of mind, and an enriched naturalism.

Cognitivism about emotions presumably arises from the observation that affects can be about something: they can be representational states, even propositional attitudes. Some scholars have attempted to reduce affects to propositional attitudes like belief or judgment, or at least to claim affects require these kind of states. In philosophy, the most common attempts at reduction of affects have generally been made for emotions, although some have also attempted to so reduce desire. Here I will criticize only theories which reduce the basic emotions to, or posit that they require, beliefs or other

propositional attitudes (for criticism of attempts to reduce desire to belief, see, for example, Lewis 1988, 1996).

<1> A note about "cognitivism"

Theories which claim emotions require or are made of beliefs have been, at least in philosophy, called "cognitive" theories of emotions. This is an unfortunate term. In contemporary cognitive science, for example, researchers freely posit mixtures of unconscious and even simple processes together with complex conscious processes into explanations of the kind of skills that would normally be called "cognitive." There is, in other words, no clear demarkation between cognitive processes and complex but noncognitive processes; rather, the only things that would clearly be noncognitive processes would be things like very simple reflexes or activities which are not neural, such as digestion. However, when confronting cognitivism about emotions, the notion of "cognitive" tends to be much more strong; and subcognitivism about emotions might correspondingly involve processes much more complex than simple reflexes.

Thus, what makes cognitive theories of emotions into "cognitive" theories is sometimes not clear for philosophers or scientists. However, although scientists have had their own debates regarding "cognitive" theories of emotion (one classic debate was held between Lazarus and Zajonc; see Scherer and Ekman 1984: 221-270), a lack of clarity is sometimes not as pressing a practical problem for the scientific study of emotions since such studies often need not be explicit about what is necessary and sufficient for a process to be cognitive. This is because if a theory posits a process which is widely

granted by other scientists to be cognitive, and the existence of the process can be demonstrated in experiments, then more conceptual clarification may be unnecessary. For example, if someone believes that emotions require appraisals which are by definition cognitive, and appraisals are granted to be demonstrated by the answers of subjects to certain questions, then the otherwise somewhat mysterious notion of an appraisal may not need further analysis for a hypothesis to be defended or a limited theory to be posed. Since my goals here require conceptual clarity, I will focus upon several theories which are cognitive in that they contain claims that one can reduce emotions to, or that emotions require, beliefs or closely related kinds of propositional attitudes. This too may suffer from serious ambiguities; for example, it could be that one kind of brain state is sometimes acting as a constituent of a propositional attitude or otherwise being used as one, and at other times it is not, so that the very idea of a state being a propositional attitude would be deceptive. However, since my goal is to criticize cognitive theories, and not to endorse them or any theory of propositional attitudes, I can avoid clearing up these ambiguities any more than is necessary to provide counter-examples; that is, often a cognitive theory of emotions is stated in a way (e.g., that emotions require particular kinds of judgments or beliefs) that can be refuted without providing more clarification about the necessary and sufficient conditions for a process to be cognitive. In this regard, we can best address cognitivism by examining philosophical theories of emotions which offer clear statements of some features of such positions.

---------- EXPLANATORY BOX 2.1 ABOUT HERE ----------

Griffiths (1997), in his criticism of some cognitivist theories of emotions, has used the term "propositional attitude theory" to describe those philosophical theories that hold that these affects require, or are, propositional attitudes. This is a useful clarification, and it also touches upon related issues which I aim to criticize (such as particular views about the role in mind of certain forms of rationality); in most of this book I will take cognitivism about emotions to be the view that the relevant affects are, are in part constituted by, or require, propositional attitudes. However, I will continue to use the term "cognitive." The primary reason is that the term is already established as a label for these propositional attitude philosophical theories. But another reason is that there are some approaches that attempt to explain affects by reference to the kind of states that we might call "high-level cognitive" states, but which are not based on propositional attitudes. In chapter 11, I will criticize the idea that emotions can be explained by symbolic models of the kind that have typified classical AI, and one might well call these kinds of theories "cognitive." Thus, I eventually aim to expand the notion of cognitive to include both propositional attitude theories and symbolic computational functionalism (and I do not claim to show that any other notion of "cognitive" is inappropriate for the basic emotions or any other affects).

I suspect that one might, eventually, find an even broader characterization of the cognitive that is demonstrably not a necessary condition for basic emotions. Thus, I am criticizing philosophical theories in this chapter, and in chapters 3 and 6, as my target cognitivist theories, but I believe it is likely that the results here will generalize to many of the "cognitive" theories of emotion held by many scientists (particularly psychologists), even though they are using the term "cognitive" in a way usually divorced

from any conception of propositional attitudes. For example, in their book on the cognitive origins of emotions, Ortony, Clore, and Collins have argued that emotions are

valenced reactions to events, agents, or objects, with their particular nature being determined by the way in which the eliciting situation is construed. (1988: 13)

If we were to understand this to be either a definition of emotion, or otherwise as a statement of the necessary conditions of any emotion, then this might be a theory which implies that emotions require mental states, like beliefs, that are propositional attitudes or at least of similar complexity. This is because the notion of how a situation is construed could require all kinds of abilities to recognize and categorize situations, to recall other situations, to draw inferences about them based upon our beliefs, and so on. Thus, if a certain form of the propositional attitude theory fails, then this reading of Ortony et al might also fail. Similar conclusions can be drawn for a host of other "cognitive" theories of emotion; I will not be undertaking a literature review of these theories, but I will be arguing for a view of emotions and of minds that is antithetical to some of the presuppositions common to some of these theories, and so it is important to recognize that the arguments against certain forms of cognitivism are meant to outline a general objection to some of these presuppositions. In this regard, it is sometimes useful (especially in building the case against the cognitive autonomy fallacy) to keep an admittedly vague contrast between "cognitive" processes which are typically involved in our conceptual abilities, especially language, and which perhaps arise primarily from neocortical neural circuits, on the one hand; and potentially "subcognitive" processes

which are typically involved in perceptuomotor control and integration, and in affect, and which perhaps arise primarily from subcortical neural circuits, on the other hand. Nonetheless, let me reiterate that the general use of "cognitive" in the sciences of mind is definitively not the defining feature of the philosophical cognitive theories of emotion that are my present targets; a cognitive theory of emotion is here understood to be a propositional attitude theory (or, in chapter 11, also a symbolic computational functionalist theory).

I will thus use the following terminology. A representation is (in the kind of organisms that are my concern here) a brain state that stands for another (not necessarily real) state or object. Representations need not be discrete (i.e., they can be magnitudes), but they must play a role in a representational system (that is, although I do not endorse holism and so do not require that each representation requires others, each representation must be part of a system that "consumes" that representation appropriately). I grant that affects utilize or are constituted in part by representations, and so have no objection to those theories (which, in some contexts, might be called "cognitive" theories) that take affects to use representations. Symbols are discrete representations which are utilized in a representational system in a way that can be properly modeled by a combinatorial syntactic system. An example of a symbolic system is a natural language. Propositional attitudes are representations of events or states of affairs, they have the special property that they are normally true or false (thus, these mental states can play a role in logical inference that a proposition can play — this is important when I discuss matters of rationality), and they are articulate representations formed of symbols or other representations. It follows then that, on this terminology, a cognitive state, or a cognitive

system, is one that requires propositional attitudes — although, as stated, I shall later weaken this to include complexes of symbols, and argue that we can also reject this weaker form of cognitivism.

In recognition that the term "cognitive" is difficult to pin down, and also that the line between cognitive and other representational processes is likely not clear (there is probably no clear line between symbols and nonsymbolic representations, for example), I will use the term "subcognitive," instead of "noncognitive" for those processes which are not propositional attitudes or complex symbolic representations. Subcognitive processes may be conscious but do not need to be so; and they can be representationally rich, but are not propositional attitudes or otherwise propositions, and they do not require language. Evidence that a kind of process is potentially subcognitive will include any of the following: it is shared by nonhuman animals of presumably simpler mental abilities (e.g., rats); it develops in humans before language and other complex cognitive abilities; it did not have to be learned; it happens or is elicited very quickly (e.g., in a few tens of milliseconds); the neural wiring that enables it is subcortical, or otherwise can operate independently of the kind of neural structures that enable abilities like language; assuming the process is a propositional attitude explains nothing more than would assuming it is a more basic representation; assuming the process is a complex of symbols explains nothing more than would assuming it is a more basic representation; the agent is unable to report accurately or at all on the process or its object or cause.

<1> Two kinds of cognitivism: reductive and doxastic

Philosophical theories which associate emotions with cognitive states like beliefs are usually of two kinds. There are theories which identify emotions with other propositional attitudes; I will call these reductive cognitive theories. There are also theories which may not identify an emotion with these other mental states, but which claim that emotions require beliefs of particular kinds. I will call these doxastic cognitive theories. The claim that the beliefs need be of certain kinds is necessary. Most of us, for example, might think that human minds must have some beliefs, and since human emotions are mental states they would require beliefs in this sense. But the doxastic cognitivist means something stronger than this; what is of importance in the doxastic theories is that emotions require beliefs which are instances of particular kinds specific to the emotion. This requirement will be made clear for each doxastic theory as needed. The reductive cognitive theories are usually trivially doxastic cognitive theories, but doxastic cognitivism need not be reductive. In the rest of this chapter, I shall show that these theories, when construed as universal claims about all emotions, are false.

Reductive cognitive theories generally have similar presuppositions, and here we can get a sufficient characterization of them by reviewing just a few of these presuppositions. Most of the reductive cognitivist theories in philosophy are of two general kinds:

(1) Judgment theories. Solomon claims that "An emotion is a judgment (or a set of judgments)" (1977: 185). Not all such judgments result in emotions, but rather "Emotions are self-involved and relatively intense evaluative judgments.... The judgments and objects that constitute our emotions are those which are especially

important to us, meaningful to us, concerning matters in which we have invested ourselves" (187). Nussbaum reconstructs, and endorses a version of, the view of the stoic Chryssipus (1987). This is the view that emotions are judgments of value, where the judgment concerns something that is essentially related to the eudaimonia — the well-being — of the subject. Nussbaum writes, an "emotion is itself identical with the full acceptance of, or recognition of, a belief" (1990: 292). This phrasing makes it seem that Nussbaum has a second order theory (where emotion is a belief about belief); but as I understand her it is the actual formation of the relevant belief which is the emotion (albeit the belief may be one we resist, and so second order epistemic matters are involved).

(2) Reduction to belief and desire. Marks proposes "that emotions are belief/desire sets... characterized by strong desire" (1982: 227), and thus "emotion reduces to belief plus strong desire" (240). Nash gives a slightly more sophisticated version of this, in a theory he calls the "new pure cognitive theory" (1989). He holds that an emotion is triggered by beliefs and desires which give rise to a dispositional state which results in a desire upon which the subject has an unusual degree of focussed attention and (potentially obsessive) overvaluation. The emotion is this state of having and focussing upon an overvalued desire (or perhaps desires).

We can, without loss of accuracy, group these views if we recognize that the cognitive element is the formation of the right kind of beliefs. There are many situations in which we form or assent to some belief (that is, in which we make a judgment) but do not feel an emotion (such as, forming beliefs unconcerned with our self-esteem or eudaimonia),

so it is clear that it is the beliefs themselves, and not the judgment, which is the operative notion in the theory. Similarly, Nash's theory begins to border upon a cognitivist theory that is not reductive, since it introduces these elements of focus and overevaluation; here I am assuming it is reductive since presumably one can be focussed upon other things and value other things without having an emotion, so that focus and valuation are not themselves just emotion under another title.

Doxastic cognitive theories are more homogenous, and can be treated as of one kind:

(3) Doxastic cognitivism. Radford (1975), Walton (1978), Shaffer (1983), and many others hold that emotions are caused by certain beliefs and desires. Shaffer explains it using an example: "I am driving around a curve and see a log across the road.... I turn pale, my heart beats faster, I feel my stomach tighten.... I slam on the brakes and stop before I hit the log. I acknowledge that when I saw the log I felt afraid" (1983: 161). His analysis of the situation is to take "an emotion to be a complex of physiological processes and sensations caused by certain beliefs and desires. Thus, seeing the log, I believed that bodily harm was likely and I desired not to be harmed" (161).

This view is quite similar to that of Marks and Nash, but the two kinds are distinct in that doxastic cognitivists take other bodily responses to be essential to the emotion — even to be the emotion — whereas Marks, Nash, and the other reductive belief/desire views are going to take the physiological response to be an unessential consequence.

<1> Empirical evidence against reductive and doxastic cognitivism

In chapters 3 and 6 I introduce arguments against both reductive and doxastic cognitivism that appeal to philosophical notions of rational action and to common platitudes about emotional behavior. These approaches are important, since philosophical differences about the import of scientific results can mean that the vast empirical evidence available to us is moot. Here, however, I will go straight to the scientific evidence, which, for at least the basic emotions, effectively demolishes both views. I will review 6 objections here.

1. The confusion of cognition with affect. One problem with reductive cognitivism is that it does not capture what is specific about affects and separates them from other cognitive states like beliefs or merely entertained ideas. For example, basic emotions are characterized by autonomic body changes (one can of course deny this, and may have to, in order to defend a reductive cognitive theory of emotion). But judgments or beliefs are not so characterized. For Marks, these body changes are themselves just features of desire: for him it turns out that an emotion is not just a combination of beliefs and desires, but of beliefs and "strong" desires. This is perhaps nothing more than a terminological difference from the kind of taxonomy we introduced here: whereas I doubt that there is anything like desire, and so I separate basic emotions from desires, Marks would group desires and basic emotions together, calling the latter "strong desires," and then construct cognitive emotions out of beliefs and the strong desires. Much the same could be said about Solomon's notion that the judgments that constitute

an emotion are "intense." But if all that was intended were a terminological change, such an approach could at best be called misleading, since: (1) there is surely something distinct about anger and desire, or anger and judgment, as these are usually understood; (2) emotions do not operate like desires are supposed to do (see chapter 3); and (3) this would make it impossible to distinguish the different emotions, since they would all be instances of a generic notion of desire; so that (4) this would make the position merely a form of doxastic cognitivism, since it would presumably be the beliefs which distinguish the emotions. Nash, on the other hand, is explicit: emotions normally have but do not need their physiological correlates. "What I deny is that bodily changes constitute being emotionally upset or perturbed, or are even necessary to such a state" (1989: 497).

Even if we avoid talk about beliefs, and instead reduce emotions to judgments, the result is similarly problematic. What separates judgments that are emotions from other kinds of judgments? The answer is the content of the judgment: the belief that is formed. In Solomon, emotions arise when we are making judgments about ourselves, the content of which matters to our sense of self-esteem. For Nussbaum's Stoic, they arise when we are making judgements about things we value. This characterization could be circular, given that much moral theory attempts to explain value — or at least valuing — in terms of emotion; if we were to accept such claims and then argue that emotions are judgments about what we value, the theory would be quite vacuous. But these theories fail on more explicit grounds. Since emotions are identified with judgments, the relevant judgments should always be accompanied by the proper emotion, and instead they are not. This has been shown by, to pick just one example, Damasio's studies of prefrontal cortex damaged patients who show no measurable loss of cognitive skills but who have, as a result of

their brain damage, emotional defects. One such subject, EVR, was studied extensively (Damasio et al 1990). This subject has an IQ of 135, and passes all the usual neuropsychological tests like a normal. But he came to the attention of the Damasio and his colleagues because he showed deficits in rational decision making. In one experiment, EVR was shown pictures of disturbing and provocative scenes. These pictures cause in normals a skin conductance response — a clear measure of the autonomic signs of affect. But EVR showed no significant response — he literally flatlined on his polygraph when he merely looked at the pictures and was not asked to describe them. This subject even reported after the test that he had noticed that he did not have the kind of feeling that he thought he ought to have for some of the pictures. EVR has the cognitive ability to recognize and describe the phenomena, but he does not usually have the appropriate emotional responses to them. Damasio's explanation of EVR's lack of reaction, and of his impaired rationality, is his own somatic marker hypothesis: Damasio argues that the bodily reaction that a normal subject has for the affect-evoking stimuli acts as a marker of that stimuli, and we sometimes depend upon this marker in making rational decisions. But regardless of whether the somatic marker hypothesis is true, EVR is a clear counter-example to reductive cognitivism, and perhaps even to doxastic cognitivism. He has intelligent, seemingly-rational judgment making abilities, makes the correct kinds of judgments, and not only has little or no affects in some of these cases, but in his everyday life performs so many irrational tasks that he is essentially disabled.

2. The inexplicability of direct neural stimulation and of abnormal cases. Other kinds of evidence of basic emotions without the kind of content as constituent or cause that cognitivism requires include the generation of basic emotions through direct stimulation of the brain by electrodes, or by what is believed to be direct stimulation from defects like epilepsy. Direct electrical stimulation of particular subcortical areas of the brain can yield affective states in humans and nonhuman animals (see King 1961, Gloor 1990, Fish et al 1993; for review see Frijda 1986: 381-86). Recall also that (as we saw in chapter 1), the neuroscientist Jaak Panksepp has argued that the basic emotions are in fact identifiable by the criterion that they can be generated by direct electrical stimulation of the brain (Panksepp 1998: 52). Also, brain damage can result in spontaneous and excessive affect. Specific emotional reactions often accompany the onset of seizures for epileptics (Ervin and Martin 1986). It has long been known that lesions in parts of the hypothalamus can cause rage in human and nonhuman animals. The classic studies of decorticate cats also first led to such observations (Cannon and Britton 1924; Bazzett and Penfield 1922; see also Bard 1928). To sustain a reductive or doxastic cognitive theory given such observations one must either deny that these are real emotions, contrary to all the behavioral evidence that is available; or somehow claim that these lesions and direct stimulation first, or at least simultaneously, generate the required beliefs of the organism. This is possible but implausible; at least, the burden of proof is surely with these cognitivists.

A related and noteworthy fact is that some emotions seem to be more easily triggered by features which are not themselves in any relevant way beliefs. For example, Zajonc has argued that failure to cool the brain properly (which can happen, for example,

if your sinuses are very severely clogged) can cause anger (Zajonc, Murphy, and Inglehart 1989). And we recognize that things like being too hot, loud noises, an uncomfortable chair, and other environmental factors can predispose us to certain emotions.

3. The problem of homology. If we accept evolutionary theory, we should expect there to be homologs of many capabilities between organisms, where more nearly "related" organisms share more common features. Thus, we should expect affects to most likely exist in other species of animals, and be more similar to our own affects as those animals are more closely related to ourselves. And we do in fact in general talk this way, and most scientific understanding of emotions has these states as being present in many species of nonhuman animals. We do not usually attribute fear to worms, but we do usefully attribute fear and a host of other emotions to cats and dogs; and many scientists readily study fear by using cats or rats or other organisms as models. Are we mistaken to do this? It would seem on a doxastic or reductive cognitive theory of emotions that we are, since presumably a cat or rat does not have the kind of cognitive capabilities necessary for an emotion on such a view. As already observed, we can weaken the sense of emotions being cognitive, so that a cat's fear is said to be merely representational. The cat is afraid of an approaching dog because undoubtedly it recognizes and categorizes the dog as a threat. But such a weakening of the requirements of what will make an emotion cognitive will fail to satisfy some of the goals of having a doxastic or reductive cognitive theory of emotion. One of the principal motivations for a doxastic or reductive cognitive theory of emotions has been to make emotions a part of

rational action by having each relevant emotion be a state with content that itself can be part of a rational "belief-desire system;" a foremost feature of this is the formation of propositions and some minimal proper logical procedures upon them (drawing inferences, expunging contradictions). Presumably mere representations, that are not part of reflectively propositional contents, do not qualify: mere representations are not true or false, for example, so cannot be consistent or inconsistent; they cannot alone play the same kind of role in an inferential system that propositions can; we cannot revise them in the same way; and so on. Similar problems arise for emotional evaluation. At the very least, doxastic or reductive cognitivism is going to have to be supplemented with a powerful theory of representation if it is going to explain how both rats and humans can have emotions that are to be reductively or doxastically construed.

Even setting aside these concerns, it seems clear that there are some nonhuman animals which are emoting and which do not have the same kinds of content as we do when we have what is purportedly the same kind of emotion. Since the state of the "fearing" cat can share many of the physiological and behavioral features that our own emotions do, we are again confronted with the question of why we would take the cognitive aspects of the emotion as more important than these other features. Taking evolution seriously suggests that the other features should be primary, such as the kind of behavioral responses (in this case, flight) shared by these animals. Finally, our growing understanding of some of the neural circuitry enabling some emotions and other affects often includes the identification of crucial roles for subcortical structures that are widely shared across mammals, and some of which may even have homologs in more distantly related species.

4. The problem of early development of the emotions. Human beings show a development of some emotional capabilities from infant (see Scherer and Ekman 1984: 73ff) to mature adult, including also the development of some affective capabilities prior to the development of our cognitive abilities. An infant can show some of the facial expressions of emotions, and after only a few weeks many of the behavioral features of some emotions — signs of anger at being frustrated, or fear when confronted when strange stimuli, or smiling when they see a mother's face. Surely such infants do not have the developed cognitive skills, however, to allow them to have the attitudes like belief and desire that a doxastic or reductive cognitive theory require (and consider also Eibl-Eibesfeldt's research discussed in chapter 1). Our best understanding of development suggests that affects like the basic emotions are capabilities that are inherited, and which can be changed by learning, including being eventually being directed or caused by propositional attitudes. This is a view contradictory to reductive or doxastic cognitivism, in which the abilities to entertain propositional attitudes of the relevant kind would have to precede the ability to have the relevant emotions.

5. The problem of neuroanatonomical differentiation. There are structural distinctions in the neuroanatomy underlying basic emotions and some other affects which are not consistent with cognitivism. This is a point well illustrated, for example, by recent research by LeDoux, who has worked to map out the neural pathways of fear and show that there is functional and anatomical separation between affective and cognitive processing systems (for an overview see 1996). LeDoux has shown that there are neural

pathways involved in fear conditioning which link to both cortical and subcortical areas. In particular, using fear potentiation studies of rats, he found that the aural cortex could be ablated and the fear-conditioned response could still be shown, working through the subcortical pathways. What was lost when the aural cortex was ablated was tone discrimination: a rat would show fear response to any tone, where before it could discriminate the tone to which it had been conditioned. In human beings there are also a host of complex pathways that operate for basic emotions, including connections between the amygdala and other subcortical structures believed to be essential to basic emotions, and also connections to various cortical areas, including polymodal and supramodal areas. The proper picture of the relation between affects and content therefore seems to be that affects can have varying degrees of cortical contribution. If any one of these cortical areas that was connected to the amygdala and other relevant subcortical structures was lost, we can expect that an affective ability could in some specific way be impaired, but that it would still remain.

The subcortical pathway that LeDoux identified for fear (and presumably such pathways could be present for other basic emotions) is similar to the kind of system that is suggested by Zajonc's research on the mere exposure effect. The affective results of these pathways are not best called cognitive, or at the very least they are surely not best identified as operating by way of generating propositional attitudes: they are faster than high-level cognition, less discriminating, and not open to report. It is also consistent with the findings of Öhman and Soares (1993), discussed in the last chapter, which provide some evidence for the theory of Seligman (1971) that some subjects are biologically predisposed for fear conditioning for some stimuli, such as snakes. Also, since Öhman

and Soares's findings were shown to be independent of lateralization, and since many cognitive functions, and especially language, are highly lateralized, this suggests that the relevant fear conditioning or recognition in question is subcortical.

6. Displacement. Finally, there is a phenomenon which has in part been studied by scientists in terms of generalization and second-order conditioning (and which may also have an analog in theories of emotional congruence in perception and attention; see Niedenthal and Kitayama 1994), and which is part of our folk preconception of emotions. It is common folk psychology that an emotion can, as it were, go searching for an object. Eric can start out angry at his landlord, and end up angry at his boss for reasons that at some other time would not make him angry at his boss. Our common understanding of such events is that Eric is in a state of anger, caused by beliefs about his landlord, and this state can begin to take different objects. But if reductive cognitivism were true, then such displacement should be impossible; instead, in having two different sets of beliefs or judgments, we would have two unrelated emotion events. And similarly for doxastic cognitivism: if an emotion requires a belief, then either we have two emotions here (because two different beliefs) or we have one emotion with two different beliefs. If the latter were the case, we could rightly start to ask about what in the emotion is shared between these two doxastic states, and this unchanging element would seem to be more essential to the emotion than the fungible beliefs that are said to be required. The former case is ruled out by the conditions of the thought experiment: we supposed that the displacement results in an emotion in cases that otherwise would not give rise to the emotion. If Eric's belief that his landlord is charging him too much money is necessary

for Eric to be angry at his landlord, how can it then be that Eric ends up angry at his boss for reasons that normally would not cause him to be angry? The anger in the latter case would seem to be better explained not by the beliefs involved, since these can sometimes fail to cause an emotion, but by some other factors. Thus, if emotional displacement of this kind occurs, it poses a counter-example to both doxastic and reductive cognitivism.

<1> Weak cognitivism

In arguing against doxastic or reductive cognitivism, I do not deny that, in humans, a basic emotion might often have some kind of belief or propositional content accompanying it; all evidence indicates that emotions in humans often are guided by propositional contents in a way that merits the title "cognitive." It is also possible that doxastic cognitivism could be true of some of the things we call "emotions;" that is, some of the things that we call "emotions" may be distinguished by reference to related beliefs, from affects which on any other scientific measures of the individual are relevantly of the same kind; such a thing might even be because social standards play a role in the concept of what that emotion is (I return to this theme at the end of chapter 4). Also, as already stated, given a weak sense of "cognitive" — so that, for example, a mental process is cognitive if it is representational — then all emotions might come to be necessarily "cognitive." Finally, of course, one is free to chose any taxonomy she desires; so we could strengthen our definition of basic emotions to make something like doxastic cognitivism true.

As we have seen, however, one reason for choosing against doxastic and reductive cognitivism is that they fail to distinguish basic emotions from other kinds of cognitive states. Our goal should be an understanding of basic emotions that is as broad and as rich as possible, and doing this requires that we look not for what is normal for, but rather for what is necessary for (or at least, most common to), the relevant emotions.

These cognitive theories are perhaps most compelling when they are used to account for those features of emotion which ally them with what would normally be called cognitive features. These include the intentionality of emotions (the fact that they are often in some sense "about" something), their evaluative nature (they are often like judgments, which can be seen as evaluations made by the subject), and their interesting connections to rationality (some see emotions as necessary to rationality, others see them as antithetical to rationality, but most see them as having a complex and significant relation to rationality). These are all features for which any theory of emotion should account, and a doxastic or reductive cognitive theory can make a quick and plausible job of this by making emotions into judgments or having them require beliefs. Beliefs are by definition intentional, they can themselves be evaluations, and on most accounts of rationality these are going to be the elements of rational thought. In chapters 5 through 8, I shall show that there are other equally plausible explanations for these features of some of the emotions — explanations which in fact have more explanatory power.

Finally, these observations are not meant to be arguments that there is no place for a cognitive theory of emotions. In fact (as I will argue again in chapter 12), if our goal is to understand the cognitive structure of emotions, then one approach should be to study emotions directly in terms of their cognitive causes and their cognitive structure. That is,

the denial of strong forms of cognitivism like doxastic and reductive cognitivism does not entail that any study of emotions in terms solely of beliefs and similar kinds of cognitive states is erroneous. Given how incomplete our present understanding of the brain and mind is, one in fact might make little or no progress in understanding the cognitive structure of emotions by any other method. Again: I do not reject the goal of understanding the cognitive structure of emotions in terms of their cognitive contents, nor even the claim that emotions are often cognitive in some robust sense; rather, it is the separate claim that the basic emotions necessarily are cognitive in a strong sense such as, for example, we find in reductive or doxastic cognitivism, that is false.

An alternative to doxastic and reductive cognitivism is a view I will call weak cognitivism: the hypothesis that the occurrent instances of relevant emotions are for humans often, but not necessarily, highly integrated with cognitive states (including propositional attitudes). This integration can include beliefs and other cognitive states causing, determining the expression, the eliciting conditions, or the intensity of, the relevant basic emotion. I endorse a form of weak cognitivism (but, as I will show in chapter 6, we need to weaken this even further by explicitly disavowing that beliefs are even normally necessary for cognitive instances of emotions). Weak cognitivism is consistent with the affect program theory.

<1> Summary: the hierarchical model of mind

If I am going to review what lessons some of the emotions can hold for the problems of intentionality, rationality, and consciousness, and for AI, it will be sufficient

to stop the taxonomic investigations here with the notion of the basic emotions. This is hardly the last word on emotions — it leaves most of those things we call "emotions" uncategorized, and it raises as many questions as answers — but it is enough to start some explorations that will reveal much about the importance of the basic emotions and the views of mind with which this understanding is consistent.

I have been concerned to describe occurrent affects, and have proposed the thesis that they are motivational states. Affects can be characterized by such properties as their duration, physiological correlates, conscious experience, behavioral correlates, and content. All of these elements play an important role in our understanding of emotions, but of these only physiological and behavioral correlates appear to be potentially sufficient to identify and distinguish an emotion. Given that there is a class of affective states which appear to be pancultural, based in inherited biological capabilities, and which are characterized by recognizable behaviors, I concluded that these are the basic emotions. These basic emotions include at least fear and anger, and probably many other affects. These are the emotions described by the affect program theory.

The affect program theory is consistent with or explains all of the objections raised in this chapter against doxastic and reductive cognitivist theories of emotion. Cognitivism is understood as the view that emotions are constituted by or otherwise require beliefs or other propositional attitudes (and subcognitive states are therefore any representational states that are not propositional attitudes); this terminology is standard to much philosophy of emotion, but not to the sciences, so we must be careful to remember that "cognitive" here is used in this strong sense. The basic emotions are clearly distinct from beliefs and other cognitive contents in a fundamental way. There is no problem that

therefore arises from those abnormal cases of spontaneous emotions, or the direct stimulation of the brain; we should expect it to be possible to stimulate the neural substrates of the affect programs directly, without having to stimulate the cognitive centers that would often be responsible for their elicitation. Nonhuman animals show these behaviors because the affect programs evolved and so likely have homologs in other related species. The development of the affect programs is also no problem. Blind children, even blind and retarded children, need not learn, but already have, these programs. The existence of subcortical emotional pathways and the extra-pyramidal enervation of affective facial expression is consistent with this, and actually suggests that it is because the affect programs of the basic emotions are phylogenetically older than our cognitive abilities that they are in part independent of these abilities. Finally, displacement of emotions is, at least potentially, explicable, since the affect program itself can have, but does not depend for its actual existence upon, a single intentional object of the relevant kind.

These findings provide us with a powerful way to view the human mind, when affects are properly accounted for: the human mind has a hierarchy of differentiable systems. These are not only modular systems, in Fodor's sense (1983; see also Griffiths 1990, and 1997: 91-97); some of them are also more fundamental in that they are required for, and constitute part of, the function of many other systems. Thus, for example, a basic emotion that has propositional content will require capabilities that themselves underlie the possibility of instances of that basic emotion without the cognitive content. Echoing Leventhal (1984), who hypothesizes that there are two distinct but parallel systems involved in affect, I can in a preliminary way illustrate the

feature of a hiearchical view of mind that is important to my goals here by making a simplified, but very useful, distinction between two gross supersystems. On the one hand, there are the subcognitive affective systems (among many other subcognitive systems, such as primary perceptuomotor control systems) which include the capabilities that constitute the basic emotions, and which can operate independently of many or most of the capabilities that typify "high cognition." In the terms of MacLean's distinctions, this would include both the "reptilian" and the limbic systems; for Leventhal, this is the emotional or affect control system. These subcognitive systems are faster than most instances of deliberative reasoning; their functioning need not be available to report (and thus are not, in this sense, necessarily conscious or cognitive); there is no reason or need to suppose that they have intentional content sufficient for propositional attitudes; and they are intimately related to homeostatic and motor control systems (such as maintaining set points in body states, and motivating actions, including the emotional actions). On the other hand, sitting (perhaps literally, in a neuroanatomical sense) above these systems are the cognitive systems, some of which may be able to operate independently of the subcognitive systems but many of which appear to need them to function properly. These are the systems which constitute the capabilities that typify "high cognition": language, the ability to plan, the ability to report on one's deliberations, and so on. Leventhal calls this the "problem control" system.

It is also tempting to assume that the affective systems are largely or wholly subcortical, and the cognitive ones are largely or wholly cortical. However, although there is perhaps some truth in this, it is not necessary to assume this, since the distinction is primarily a functional one; and even some phylogenetically ancient functions have

been "rewired" in primates to involve neocortical structures, and so the functional notion of subcognitive capabilities need not correpond to this basic anatomical distinction.

This two-tier distinction is too simple: a mature science of mind will find it more useful to refer to many systems, not easily grouped into two sets, but nonetheless clearly hierarchically arranged. However, even roughly hewn into two groups, the hierarchical view of mind is useful for drawing out a number of issues. First, it points us towards a very different way of thinking about mind, and therefore a very different kind of theory of mind, than is typical to contemporary philosophy, where critical issues are often framed in relation to propositional contents, or lack thereof. The basic emotions, and many other affects, are clearly able to operate independently of many cognitive skills, and the neural circuits that constitute some of them appear to be centered in subcortical regions or in brain structures that are functionally independent of the kind of abilities that enable propositional attitudes. Furthermore, our evolutionary understanding of the basic emotions is encouraged by the observation that other mammals, which share with us strikingly similar subcortical anatomies, also exhibit many of the same affects, including some of the basic emotions. This is all consistent with a bottom-up view of mind, in which affects and perceptuomotor abilities are understood to be phylogenetically and functionally prior to, and likely necessary for, cognition.

Second, this simplified perspective on the hierarchical view of mind also helps us to clarify where disagreements about the taxonomy of affects are, and are not, substantial. There is in fact a great deal of implicit agreement among many scientists for the essential features of the affect program theory. Disagreements tend to arise about how much we need to add to get a full-bodied "emotion." Roughly, and using again the simplistic two-

tier idealization, it may be that for some basic emotions we could outline two kinds of definitions, or identity criteria. The first, of the kind I utilize here, would refer primarily to the subcognitive systems to identify the capabilities and neural circuits that constitute the basic emotions. It would expect the exercise of those emotions not to require the kind of cognitive skills that are special to humans, since homologs of these emotions exist in other animals. The second kind of definition would refer also to cognitive systems, and thereby make use of a broad, or "thick," notion of the basic emotions, perhaps construing them as necessarily conscious, or necessarily propositional attitudes. Which kind of definition one should use is not an issue we need spend much time debating; I have argued that something quite like the former is a richer notion, which avoids the fundamental confusions encouraged by the latter. But the latter notion is wholly consistent with the substantive claims made throughout this book, as long as it is recognized that affective systems that are not necessarily propositional attitudes are themselves necessary to the emotion in the thick sense, and sufficient for it in the sense the affect program theory uses. Given this, I hold no disagreements with anyone who accepts that the kind of things that happen in the affect program theory are necessary to the relevant basic emotion, but then defines that emotion in a cognitive way or even a necessarily social way. Disagreements arise, instead, with those who either (1) deny that the subcognitive elements on the hierarchy are necessary, or (2) define the basic emotions as cognitive and then use such a definition in too-general a way. (The first disagreement is what we saw in reductive cognitivism: the view that the beliefs and other kinds of cognitive states are alone necessary, and the other features picked out by the affect program theory are unnecessary.) Given this understanding, a very great deal of

agreement should be possible between what is said here and the majority of views on the relevant emotions.

[i] The term "basic" has sometimes been associated with the view that all emotions are constructed out of some combination of the basic emotions. I do not endorse this view. However, I continue to use the term because all the alternative terms are traditionally even more loaded: "primitive," "fundamental," "innate," and so on are also potentially deceptive.

[ii] I consider a functional account (at least in the sense I use the term here) to be consistent with a type-reduction account.

[iii] See (Niedenthal et al 1994a) and (Niedenthal et al 1997) for an example, explanation, and criticism of the problems that can result from the application of such models of emotion as they relate to emotional congruence in perception. Their adoption of a categorical model of emotions is consistent with the affect program theory as opposed to these one or multi-dimensional appraisal theories and related theories.

[iv] My concern, in making this distinction, is not with the role that context-dependency may play (for example, a sense of "disposition" that may concern a neuroscientist), but rather to clarify an ambiguity in the use of emotion terms. As I also explain below, this is closely related to an issue about whether our use of emotion terms entails that there is a measurable body state or rather is just a way of talking about the likelihood of certain kinds of behavior (and which may or may not require such a measurable body state that constitutes a motivation during, or just before, the occurrence of that behavior).

[v] I will not discuss temperament in this book. It is a difficult and interesting topic. See (Steinmetz 1994).

[vi] As discussed in chapter 3, the interpretationists can be said to hold a position like this, as long as the contrary notion of there being a correspondence between desire and an actual body state is understood sufficiently strongly (as it must be for any realistic naturalism about affects and their role in mind). Similar issues arise for various defenses of internalism, discussed in chapter 8.

[vii] These distinguishing features are tilted towards philosophers in that they include relation to cognitive content — and, in particular, the question of whether an affect is a propositional attitude or otherwise a similar kind of complex cognitive state. But these features are similar to those used by many scientists, such as in Frijda (1986; see 1-4).

[viii] This is largely my failing; I am not familiar with this literature, and also do not know how to respond to the many objections that it can raise. E.g.: a social constructionist might hold that some emotions are as long or short as they are because this is what is considered appropriate in that culture. Or some

differences in duration could be products of other defining features of emotions and as such secondary in importance to those features. And so on.

[ix] I will not discuss this experiment which is often taken to show autonomic responses are insufficiently complex to specify emotions. A number of criticisms have effectively shown that this experiment does not establish this; see Damasio (1994), Gordon (1987), Griffiths (1997), LeDoux (1996), Levenson (1992).

[x] For example, in Ekman, Levenson, and Friesen (1983) both happiness and anger were found to result in an increase in heart rate and in temperature. Although both increased dramatically more for anger than for happiness, one might question whether weak anger would appear like an intense happiness on these two measures alone.

[xi] My target here is to get at a notion like Chalmers's notion of *psychological consciousness* (1996: 25-26). I prefer the term "working" since it separates the concept from any specific body of theory. My criterion here that the state be reportable is much stronger than Chalmers would require, however; I will weaken this in chapter 9. In DeLancey (1995) I used the term "functional consciousness" for this; but I use "working" here to avoid possible confusion with teleofunctional notions.

[xii] Ben-Zeev argues that "It is meaningless to say that an agent is unaware of, or misidentifies, his feelings"; instead, "An unconscious emotion then is usually one about whose nature the agent is not clear, but is aware of many of its components" (1987: 401). There is therefore an interesting question about whether this is correct, and so alexithymia is a matter not of failing to identify feelings, but rather failure to identify the affect from which they arise. I shall remain agnostic about this; I take it to be an empirical question — although there may be a conceptual issue in clarifying what "identifying one's feelings" is. Here, we need only the weaker case of failing to recognize the emotion as what it is.

[xiii] For a dated but useful review of related research of nonhuman primate facial expression, see also Chevalier-Skolnikoff 1973.

[xiv] Stating that the action is a consequence of the program can be a little deceptive, since the action is not separable from the program; a car's engine can idle while the car sits still (the analogous case to not acting on the program) but the activity of speeding along in the car cannot be separated from the running of the engine (the analogous case to the action program being uninhibited and leading to action).

[xv] But there remain substantive empirical and conceptual issues about what the relevant inhibition is. It may be that a combination of things, some best called

inhibition, others best called redirection, others disconnection, are involved in an occurrent action program not resulting in emotional action. There is also a conceptual issue about whether these can ultimately be distinguished in a robust way. Here I will not hypothesize about which, if any, of these alterations of emotional function is operating, and will use "inhibition" as a broad term to cover all of them.

[xvi] I believe that the arguments that I will make that refer to type-specific function and eliciting conditions (these arguments occur in chapters 5 through 8) could in fact be made with the weaker supposition merely that there are some type-specific functions and eliciting conditions. However, such arguments would be convoluted and lack any intuitive appeal, and I believe that these hypotheses regarding function and eliciting conditions are surely close enough to the truth to be appropriate.