

Chapter 3: interpretationism

In this chapter I want to accomplish two things. First, to give an example of how the failure of cognitivism about emotions can be important for contemporary theories of mind. Second, to argue against a potential form of cognitivism that is resistant to the criticisms raised in chapter 2. There, we learned that much scientific evidence is not consistent with a doxastic or reductive cognitivist theory of the basic emotions. However, this kind of evidence is only going to be compelling if our theory of mind is itself open to these scientific results. But there are a number of theories of mind that have been extremely influential and for which this scientific evidence would presumably not apply. The most prominent of these are the various interpretationist views.

Interpretationism and functionalism have been the most favored alternatives to the kind of simple reductive naturalism that has long been out of favor in philosophy; in chapter 11 I discuss one form of functionalism, and in chapter 13 I will briefly discuss why I believe the conception of naturalism shared by much contemporary philosophy of mind is erroneous. In this chapter, I will consider interpretationism, and draw out its relation to basic emotions by considering some emotional actions. These emotional actions raise special problems for theories of mind and action. Ultimately, I will show the affect program theory better explains these actions. This advantage in explaining behavior is evidence in favor of the affect program theory as the best account of the nature of the relevant emotions. By way of contrast, I will show that these leading irrealist theories of mind fail to adequately explain these actions. This

provides strong evidence that these theories fail to account for the nature of these emotions. I believe that it also shows that these theories are not appropriate theories of mind.

----- BOX 3.1 ABOUT HERE -----

Because interpretationism takes cognitive states to be ascribed, it entails an irrealism, and in practice a reductive cognitivism, about emotion which it can be particularly difficult for a naturalist to answer. This is because the scientific evidence can often be pushed aside since it supposedly tells us nothing about the mental states which the interpretationist claims we ascribe (and where an emotion is either a new one of these states, or reducible to some of these). My strategy in this chapter therefore is to build the case for a naturalist theory of emotions by confronting what I take to be the most resistant alternative (interpretationism) by primarily using conceptual arguments and common-sense examples (both of which should be admissible by the interpretationist).

First, I will introduce a special class of emotional actions. Then, I will discuss Dennett's intentional systems theory, and show that Dennett's irrealist position is unable to account for these actions. Considering both these actions and other problems, I will then discuss the more difficult case of the interpretationism of Donald Davidson. Throughout, I will show how their respective versions of interpretationism about belief and desire requires that they be reductive cognitivists about emotion, and I shall construct explicitly interpretationist theories of emotion within their frameworks. In conclusion, I will return to some of the relevant scientific evidence for the affect program

theory, and show how that theory is best suited to account for the actions introduced in this chapter.

----- BOX 3.2 ABOUT HERE -----

<1> Emotional actions

In this chapter, I will use “emotional actions” as a non-technical term to label those actions which can be explained (not necessarily exclusively) by citing a basic emotion as their cause. Emotional actions are most clearly identified when contrasted with a belief-desire account of action. In such an account, an intentional action is explicable via the beliefs and desires supposedly held by a minimally rational agent. This is a useful starting point because the interpretationists assume that our mental states are sufficiently explained via belief and desire. In contrast, emotional actions are normally described as being the product of emotions.

We can identify at least three ways in which the common-sense explanation of an action will relate to a belief-desire explanation of the same action by distinguishing between four kinds of emotional actions. I call these kinds of actions B-D consistent actions, B-D inconsistent actions, B-D independent actions, and B-D post-functional actions.

(1) An emotional action is *B-D consistent* just in case the action is what would be predicted by a belief-desire account.

Suppose that Eric runs away from a rabid dog. It may be true both that Eric ran because he was afraid of the rabid dog;¹ and also that Eric ran away because he believed the rabid dog would bite him and he desired not to be bitten. These are the kinds of cases in which emotion poses no challenge to belief-desire psychology; such cases also provide the most plausible supporting evidence for the reduction of emotion to belief and desire (e.g., Marks 1982).

(2) An emotional action is *B-D inconsistent* just in case the action is inconsistent with what would be predicted by a belief-desire account.

These are actions which not only would not be predicted or posited on the belief-desire account, but they are either of a kind that would be predicted or posited *not* to occur, or are of a kind that is incompatible with the kind of action predicted or posited. An example could include: Karen is afraid of dentists and so avoids them, although she believes that a dentist will do more help than harm to her teeth, and she desires that her teeth be helped, and so on.¹¹

B-D consistent emotional actions pose no special problem for the irrealist position about emotion if the irrealist is going to depend upon a belief-desire theory of action. This is not so with B-D inconsistent emotional actions. These actions are counter-examples to the view that emotions are reducible to belief and desire, and provide some strong evidence in favor of a naturalist theory of emotions. However, B-D inconsistent actions (at least of the kind of which I am aware) make problematic counterexamples for a number of reasons -- for example, some deny that such actions are even possible. We can therefore make

the case against interpretationism more simply by focussing upon specific cases of B-D independent emotional actions.

(3) An emotional action is *B-D independent* just in case there is no plausible B-D account of that action.

The most obvious such actions are purely expressive actions, such as kicking a tree when mad at your boss, or jumping for joy on some good news. Such actions simply have no plausible B-D account. However, these expressive actions are open to the criticism that they are automatic behaviors, like digestion or a reflex jerk of the knee, and thus of no interest to a B-D account. We need to focus upon a special kind of B-D independent action to sidestep this moot point: B-D post-functional actions.

(4) An emotional action is *B-D post-functional* just in case the action continues beyond the satisfaction of any of the agent's relevant desires.

Although not inconsistent with a belief-desire account, these actions result in an explanatory gap for such an account. Consider the following plausible story of a B-D post-functional action:

Case A: Eric is terrified of the rabid dog that is before him. He also believes that it will bite him, he desires not to be bitten by it, and he believes that by running into his house and closing the door he will not be bitten (because he knows that the dog cannot get through the door, cannot

turn a door knob, etc.). These are all of Eric's relevant beliefs and desires.ⁱⁱⁱ Eric runs into the house, frantic; once inside, he not only locks the door, but Eric then runs upstairs and into another room. He closes and locks that door, and then he moves into a corner: the place in the room farthest from where he believes the dog is.

I propose that the case as supposed is possible — indeed, that there is nothing extraordinary about such a case (especially if we imagine that Eric is young, or, better yet, the kind of person we might call fearful). But the action of Eric *after* he is already in the house with the door closed now continues beyond the scope of the belief-desire explanation of his action. His desire not to be bitten was satisfied when he got inside and closed the door. From the perspective of the belief-desire account, everything else that he does is extraneous. I have set up this case so that the action of closing and locking the second door is plausibly a continuous extension of, and of a kind with, the closing of the first door; so that there is one continuous action that extends beyond the realization of the relevant desire.

We can imagine any number of such cases, involving different emotions. Consider the following supposition:

Case B: Tim is furious at a rabid dog that has attacked his daughter. He desires to kill the dog, and he believes that by shooting it with his gun he can do so (again, these are all of his relevant beliefs and desires). He decides to shoot it, hurriedly finds and loads his gun, chooses the best

approach to the dog, sneaks up on the growling canine, and shoots it. On the first shot the dog falls down, obviously dead (and Tim at this moment knows that it is dead). Tim fires all the rest of the rounds in the gun into the body of the dog.

How can we understand Tim's action? To most of us, there is nothing mysterious about it: Tim was enraged, and his behavior would be explained, even predicted, given that observation. But by supposition his desires were satisfied with the first shot. Again, his continuation of the action of firing is, from the perspective of a belief-desire account, post-functional.

<1> Intentional Systems Theory

Post-functional emotional actions pose a special problem for the interpretationist theory of mind. There are many types of interpretationism (see Dennett 1993d for an overview); I will focus upon Dennett's intentional systems theory as the most basic example. Intentional systems theory is irrealist because it maintains that intentional states are the products of interpretation. They are ascribed to an agent by an interpreter but are not natural states to be found and measured by natural sciences. Intentional terms like "desire" therefore are used for the interpretation of behavior and are not genuine scientific description of the mental.

To use intentional systems theory to explain behavior and mental states, we understand actions of a complex system by attributing to it beliefs and desires, and then making predictions based upon those attributions. The

procedure is to (Dennett 1993a: 17):

- (1) treat the object whose behavior is to be predicted as a rational agent;
- (2) figure out what beliefs the object ought to have; and
- (3) figure out what desires the object ought to have.
- (4) Assume the object reasons through (2) and (3) (via the rational abilities granted by (1)), and derive a course of action for the object; this is the course of action that we as observers predict.

This is the *intentional strategy*, and looking at something in this way is to take towards it the *intentional stance*. If we see a human being (the object) in a room that catches fire, we might attribute to him the desire to survive; and attribute to him the beliefs that the room is on fire, that fire can kill him, and that fleeing the fire would prevent the fire from killing him; and finally predict that he will run out of the room to flee the fire. If he does run out of the room, we have made a successful prediction, and we take this as showing the object is rational and has these beliefs and desires, and actually did undertake the predicted action.

Prediction plays the role of making this a theory that is, in some sense, appropriately scientific because falsifiable. Without this, intentional systems theory is nothing more than an attribution of actions and mental states, and is unenlightening, since any behavior can be labelled as the product of any of endless different combinations of beliefs and desires.

One of the features of this strategy is that the intentional states, and the actions these states are to explain, are tightly connected. A belief is the kind of thing that explains why some action followed some desire (our agent believed

the room was on fire, since he fled from the room and desired to avoid being killed by the fire); a desire explains why some action followed some belief (our agent desires to survive, since he believed the room was on fire, believed that fire could kill him, believed fleeing the room would prevent that, and fled the room); and what makes a behavior one kind of action rather than another is the beliefs and desires attributed to the acting object (our agent fled the room — as opposed to running out into the hall in order to, say, get home more quickly — because his desire was to survive, and he believed the room was on fire, believed that fire could kill him, and believed that fleeing the room would prevent that). Being a rational agent is at least making the appropriate inferences from ascribed beliefs in order to satisfy desires. We can thus also use this stance to understand actions that have occurred, and to understand what a belief or a desire is, and all of this is comprised by *intentional systems theory*. To quote Dennett on belief: “*all there is to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence all there is to really and truly believing that p (for any proposition p) is being an intentional system for which p occurs as a belief in the best (most predictive) interpretation*” (1993a: 29). The theory entails the same thing for desires: all there is to being a true desirer is to be a system whose behavior is reliably predictable via the intentional strategy when the desire in question is so ascribed and this ascription yields the most predictive interpretation.

In this scheme, emotion must fit as some form or combination of beliefs or desires. Dennett has not discussed emotion in his works except to sometimes mention them as intentional states to be ascribed via intentional systems theory (e.g., 1971: 87), but in fact his own explication of intentional systems theory

requires a reductive cognitivist theory of emotions. This is because the intentional stance aims to explain intentional behavior by ascribing beliefs and desires, and, in this way, provide a reduction of folk psychology. Here we must recognize that in intentional systems theory, “belief” and “desire” are not merely the folk psychological terms (and so should not be taken to be *prima facie* distinct from emotional terms) but rather are derived, technical terms. When he gives intentional systems theory in its most general form, Dennett writes that

One predicts behavior [from the intentional stance] by ascribing to the system *the possession of certain information* and supposing it to be *directed by certain goals*, and then working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions. It is a small step to calling the information possessed the computer’s *beliefs*, its goals and subgoals its *desires*. (1971: 90)

Like some forms of decision theory and game theory — which Dennett recognizes as intentional systems theory’s “close kin” (1993b: 58) — the theory takes these two things as the sole primitives and aims to analyze all actions with them. Emotional actions, therefore, have to be so explained. The intentional systems theorist has left himself no choice but to be a reductive cognitivist about emotion.

We are now in a position to consider whether intentional systems theory can explain the post-functional emotional actions. On an intentional systems theory account of case B, we must first ascribe to Tim the beliefs and desires he ought to have. Presumably the most important of these — the beliefs and desires

relevant in this case — are Tim's belief that the dog is rabid, that it attacked his daughter, that it might attack Tim or someone else, that killing it would render it unable to attack another; the desires include wanting not to be hurt by the dog, wanting his daughter not to be hurt by the dog, and to kill the dog. But on these most plausible ascriptions of beliefs and desires, Tim's action after the first shot is absurd.

We can attribute to Tim further desires like the desire to hurt the dog, but even assuming we can justify the claim that this is the kind of desire that Tim ought to have, Tim presumably knows that the dead feel no pain, and therefore that the additional shots are not satisfying such a desire. Similarly, we can attribute to Tim the desire to shoot the dog until he is out of ammunition, or to shoot the dog when it is dead, and so on. But Dennett wants his belief-desire psychology to be more than a vacuous interpretation, so he needs there to be something more robust to desire than just a relation between ascribed actions and beliefs. Recall that the intentional stance is supposed to yield predictions, and before we make the prediction we are to ascribe to the agent those desires which it should have. But the most plausible desires, the desires that Tim ought to have, and those that would be consistent with the other desires that Tim has (the desire to secure the safety of his daughter, etc.) are not desires like emptying the magazine into the dog regardless of whether it is already dead. There is no reason, given that Tim believes dead dogs feel no pain, that they don't attack people, and so on, that Tim ought to desire to shoot the dog when it is already dead. This is not to deny that desires (should there be a general motivation state which deserves this nomination) may accompany or be the product of an emotion; there may even be a sense in which some understand desire such that

Tim does desire to shoot the dead dog. But if the emotion causes or is accompanied by another desire, a desire which falls outside the usual ascription of desires the agent should have, then we have not reduced emotion but rather introduced it as an additional thing.

If we suppose that Tim's desires are those listed above, then the action is post-functional: there is a continuity between what happens up to and including the first shot of the rifle, and what happens after this. This is one action, and it continues beyond the satisfaction of the most plausibly attributed desires. Since the action continues after the satisfaction of these desires, the whole action is unpredictable on the intentional stance. This claim that all the shooting is one continuous action is an important one: it is meant to stand in denial of the possible claim that the intentional stance explains everything up to and including the first shot, but then what follows is irrational. Such a move would be an *ad hoc* division of actions into the rational and irrational based on their predictability; I elaborate this objection below.

Case A poses the same kind of problems for intentional systems theory. By supposition, Eric knows that getting past the first door will prove sufficient impediment to the dog, were it to attack him. His additional behavior is inexplicable from any credible attribution of beliefs and desires, and hence would not be predictable from the intentional stance. The relevant desires that Eric can reasonably be said to have or which he should have — in this case, primarily to get safely distant from the dog — are satisfied once he is through the door. Running upstairs and closing another door is a continuation of the action and this continuation cannot be said to satisfy that desire or any other desire that could be reasonably attributed on the intentional stance account.

Intentional systems theory is not meant to be an analysis alone but a predictive framework. To rephrase these cases in the form of an objection to the intentional stance, then, is to recognize that from the intentional stance B-D post-functional emotional actions will not be predicted. The intentional stance predictions depend upon some action being expected to satisfy a desire of the agent, given beliefs held by the agent. We know, however, that B-D post-functional emotional actions like those in the examples above go beyond the satisfaction of the relevant desires that can plausibly be attributed via the intentional stance — the desire to kill the rabid dog, or to get away from the dog — so that on the theory the B-D post-functional emotional action either should not occur, or will at least be unpredictable. We also know that these individual tales are not unique, and that emotional behaviors are common behaviors and we recognize actions like those in the cases above as being possible instances of significant types of human behavior. This shows that the intentional stance fails to predict, and so explain, the B-D post-functional emotional actions, and therefore is *prima facie* inferior to any theory of the relevant emotions that does explain these kinds of actions.

But there are some potential objections the intentional systems theorist can raise, and I will consider these now. The goal of intentional systems theory is to provide a reduction of mental states to a more manageable domain: “the claim that every mental phenomenon alluded to in folk psychology is *intentional-stance-characterizable* would, if true, provide a reduction of the mental... to a clearly defined domain of entities whose principles of organization are familiar, relatively formal and systematic” (1993b: 68). Presumably the intentional systems theory reduction does not need to be total; rather, as Dennett aptly puts

it, “A prospect worth exploring... is that folk psychology (more precisely, the part of folk psychology worth caring about) reduces — conceptually — to intentional system theory” (1993b: 66). Granting that emotions of the kind we describe here are worth caring about, the intentional systems theorist might still object that the B-D post-functional actions are themselves not significant. One might deny that the emotional post-functional actions are actions, and instead call them mere behaviors — something like a blink or a sneeze, although admittedly more complicated. What reason could be given to call them mere behaviors? The most likely course might be to say that they are unintentional or that they are arational (outside the scope of rationality, just as is digestion or blood flow) or irrational (contrary to rationality).

Although intentional behaviors are clearly actions (non-philosophers need to note that, in this context, “action” is a technical term for philosophers, meaning something like events caused by a thinking agent as a consequence of — not necessarily conscious — thoughts, usually best described as propositional attitudes), on some definitions of “intention” there can be actions that are not intentional. The notion of intentional or voluntary action is very problematic,¹⁰ and different intuitions and working definitions abound. However, one philosophical standard is to hold that intentional actions are those actions that fulfill a plan. This standard is rather strong, and therefore it is safe to take it as at least sufficient for an action to be intentional.

The examples used above were special in that the actions undertaken accomplish a goal, but the actions then continue beyond the realization of that goal; part of the behavior is post-functional, from the point of view of the most

plausible belief-desire account. Indeed, there are other kinds of emotional behavior, such as B-D inconsistent behaviors, that we regularly see, but which so clearly fail to satisfy any plausible desire that they are open to the challenge that they are wholly irrational. There are also emotional actions which are purely expressive, such as if Karen were to kick a tree because she was mad at her boss, or if Adam pulls his hair because of his grief over some loved-one's death. These kinds of behaviors also pose a problem for a belief-desire account because there is no belief-desire explanation of such actions; but since they might be said not to accomplish a specific goal on the belief-desire attributions, they might be dismissed as unintentional behaviors. Such mere behaviors could be shouldered off as only needing explanation from a "low-level," non-intentional design-stance view. I don't think that this strategy will work, but these cases require a much more complex defense. In contrast, the post-functional actions of the kind given in the examples are themselves actions because they accomplish a goal, and do so by exploiting and fulfilling a plan (finding the gun, loading it, choosing and utilizing the best approach for attack; or fleeing on the most direct course to the door). And again, these cases are plausible and common enough in form for us to generalize that they represent instances of common kinds of behavior. As such, the kind would be intentional, B-D post-functional emotional actions.

Another objection might be that emotional actions are arational or irrational; that the action was rational until it accomplished its goal, then became otherwise as it continued past that accomplishment. It could thus be argued that intentional system theory is meant to predict only rational action, and so should not be expected to account for emotional actions. But this line of argument

would also beg the question. Dennett rightly observes that it is difficult to get a precise characterization of what rationality is; and that, nonetheless, many of us still proceed to make claims about rationality, or to use it in our theories (1993c: 98). But whereas it may be permissible to use an intuitive notion of rationality either as a starting point for trying to understand rationality or as an evaluation of behaviors as rational or irrational, Dennett is using his intuitive notion of rationality as part of a predictive theory in which it plays a crucial role. This would not matter so much if we were able to assess the predictions of intentional systems theory, since successful predictions could provide us with some reason to believe that the notion of rationality in use was adequate. But, it is difficult to assess the explanatory or predictive success people could achieve by using intentional systems theory, because no one has yet conducted the relevant experiments. And even *prima facie*, it is not clear to what extent we actually do predict each other's behavior by taking the intentional stance. Dennett offers examples of successful prediction of each other's actions; for example, when we drive our cars and successfully negotiate traffic by in part predicting the actions of the other drivers. But it is not obvious that such cases are typical cases of how we predict behavior (as opposed to being chosen because it fits the theory). Nor is it clear that our ordinary predictive success in such cases really depends upon our assuming that other drivers are rational, and ascribing beliefs and desires to them, as Dennett claims we do (for example, we may also ascribe emotions to them, based upon their facial expressions, or posit that they are drunk, based upon the way they are driving, and so on). So we have no reason to believe that intentional systems theory's working and admittedly imprecise notion of rationality is providing useful predictions, and so is an adequate working notion

of rationality. Unless there is some principled way to single out emotional actions as, in contrast, arational or irrational, to come along after the fact and call emotional actions arational or irrational behaviors would be *ad hoc*.

Thus, these B-D post-functional emotional actions should count as actions of the same significance as the other kinds supposedly predicted by the intentional stance. These actions are not directly a problem for any intentional stance predictions themselves (except when they contradict or exclude those predictions) since they are not about to be predicted. However, the claim that the intentional stance explains all or a significant portion of our mental lives is a separate claim. The stories about Eric and Tim are not incredible, and if we were there and knew something about the situations we might very well have predicted — where the intentional stance cannot — that Tim would shoot the dog several times, or that Eric would run farther than necessary to protect himself from the dog. These kinds of action fall outside the explanatory ability of intentional systems theory, but well within the domain of folk psychology. We have thus observed that there are significant kinds of emotional actions that the intentional system theorist will fail to predict, and that therefore the theory cannot reduce or explain the emotions which caused these actions. Another kind of explanation is needed if we are to account for these emotions and the actions that they can motivate.

<1> Davidson's Interpretationism

B-D post-functional emotional actions raise problems for Davidson's interpretationism which are similar to those we have seen for Dennett's

intentional systems theory. However, the case is slightly more difficult to formulate, since Davidson's action theory does not use prediction as a justification mechanism. What we can do, however, is show that there are problems with Davidsonian interpretationism which should raise our doubts that his analysis of the relevant mental states is successful; and, more importantly, we can find a better alternative. In this section I will take the first tack; in the next, the latter.

For Davidson, states like belief and desire are ascribed to agents when we radically interpret their activities and language. When confronted with an agent I want to radically interpret, I start by assuming that most of what it believes must be true, and work to create an interpretation which makes it maximally consistent both with itself and with my own language and beliefs (Davidson is not claiming that we actually ever are in the situation of radically interpreting anyone; rather, the notion is a conceptual tool, a kind of description of the nature of meaning and action at an idealized limit). Actions, in turn, are both rationalized, and said to be caused, by *primary reasons*. Primary reasons are pairs of beliefs and *pro-attitudes*. For Davidson, pro-attitudes are essentially what would normally be called "desires" or "wants;" nothing more is needed to explain actions:

Fortunately, it is not necessary to classify and analyse the many varieties of emotions, sentiments, moods, motives, passions, and hungers whose mention may answer the question, "Why did you do it?" in order to see how, when such mention rationalizes the action, a primary reason is involved. Claustrophobia gives a man's reason for leaving a cocktail

party because we know people want to avoid, escape from, be safe from, put distance between themselves and what they fear. Jealousy is the motive in a poisoning because, among other things, the poisoner believes his action will harm his rival, removes the cause of his agony, or redress an injustice, and these are the sorts of things a jealous man wants to do. (1963: 689)

Thus, like the intentional stance theorist, the Davidsonian interpretationist only has recourse to belief and desire to explain actions, including emotional actions.

This is borne out by Davidson's analysis of pride. In reconstructing Hume's view of pride, Davidson's endorses a judgmentalist theory of emotion in which affects play no essential role. Recall that judgmentalism (which, as we already saw, in slightly different forms has been advocated by Solomon 1977 and Nussbaum 1987, 1990) is the view that an emotion just is a kind of judgment that includes or contains some evaluative element. Thus Davidson writes that "The theory that I have constructed identifies the state someone is in if he is proud that *p* with his having the attitude of approving of himself because of *p*, and this in turn (following Hume) I have not distinguished from judging or holding that one is praiseworthy because of *p*" (1976: 753). What is of interest is that Davidson, in revising Hume, rejects that part of Hume most consistent with the naturalist view. Hume recognized that beliefs could not alone motivate and that some affective passion must be involved for there to be a rousing to action. But Davidson rejects the necessity of such an affective element:

Hume surely did often, and characteristically, assert that a pleasant

feeling, or a feeling of pleasure of a certain sort, was essential to pride, whereas no such feeling is essential; and, more important, such an element does not help in analysing an attitude of approval, or judgement. (1976: 754)

Davidson also claims that “what Hume called the passion has no place in the pattern” of elements Hume used to explain the nature of passions and the actions they motivate (1976: 754). Thus, the evaluative element of pride is not essentially affective for Davidson except in that it may require some pro-attitude(s), and so there is nothing distinguishing such an emotion as a kind from other kinds of intentional states like belief and desire, except their particular logical form. Judgmentalism, for the Davidsonian interpretationist, is a propositional-attitude based cognitivist theory of emotion. It is clearly at least doxastic cognitivism, since for Davidson “belief is central to all kinds of thoughts” (1984: 156), including fears and presumably therefore all other basic emotions. The view is technically not reductive cognitivism, but the difference is merely formal; since the pro-attitudes are essentially just desires or wants, this cognitivism is very close to being a species of belief-desire reductivism, where the evaluative dimension of an evaluative judgment can be glossed as a desire of some kind. For example, being afraid that the rabid dog will bite you will amount to something like a belief that the dog is rabid, that it could bite you, and the evaluative judgment that being bitten would be very unpleasant or bad; this evaluative judgment in turn amounts to nothing more than the belief that being bitten would cause pain or scarring or etc., and the desire to avoid pain or scarring or etc. What separates Davidson from the belief-desire reductionist is

only his claim that emotions will have logical structures not reducible to the logical structures of belief and desire (1984: 156); otherwise, the failure of such a reduction lies not in any differences in the relevant kinds of natural states, such as the physiological changes accompanying an emotion, since these are inessential to the emotion.

Davidsonian interpretationism faces the same problem as does the intentional systems theorist when attempting to explain a post-functional emotional action. Surely Eric and Tim can avow to their beliefs and desires; but then we have again the problem of explaining, in a belief-desire account, why they act as they do. Davidsonian interpretationism stumbles over the post-functional actions for the very same reasons we saw that intentional systems theory failed: the B-D post-functional actions continue beyond the satisfaction of the pro-attitudes the agents could reasonably be ascribed.

Ultimately, Davidsonian interpretationism is inadequate to the task of explaining these actions because the elements of folk psychology which interpretationism aims to save are insufficient to explain some emotional actions. We were able to locate the failure of intentional systems theory by observing that it fails to predict and so explain these kinds of emotional actions. For Davidsonian interpretationism, explanation does not rely upon prediction; instead, we can locate the failure of the explanation in Davidson's posit that the agent interpreted is rational in a special sense. This rationality basically amounts to an idealized ability to reason logically, which we might think of as an application of decision theory:

The cogency of teleological explanation rests... on its ability to discover a

coherent pattern in the behaviour of an agent. Coherence here includes the idea of rationality both in the sense that the action to be explained must be reasonable in the light of the assigned desires and beliefs, but also in the sense that the assigned desires and beliefs must fit with one another. (1984: 159)

And:

to the extent that we can see the actions of an agent as falling into a consistent (rational) pattern of a certain sort, we can explain those actions in terms of a system of quantified beliefs and desires. (1984: 160)

Presumably, to be rational, the agent must be able to draw some inferences from its belief set, and some inferences regarding its desires, sufficient for it to be able to act in order to sometimes satisfy some desires; it must also demonstrate some consistency in its beliefs. Let us call this B-D rationality. To be B-D rational is to satisfy, to some degree, standards regarding one's beliefs and one's actions in light of one's desires and those beliefs. Such standards are usually stated as ideals: being fully consistent, making all the useful inferences, taking all the actions that will satisfy one's desires given one's beliefs, and so on (see Cherniak 1986 for an overview and criticism). The problem which some emotions pose for Davidson's interpretationism — including the emotions that motivate B-D post-functional actions — is that these emotions are normally attributed to the agent even if they result in actions which fail to be B-D rational.

It seems unlikely that Davidson could revise his interpretationism to

include, along with pro-attitudes, the appropriate kinds of emotions — that is, emotions which can motivate B-D post-functional actions. First, as we have observed, if the radical interpreter adds emotion to the interpretive tool kit, and then claims that emotions motivated these actions, she must explain in what sense emotions could motivate the actions. But Davidson provides no other tools but pro-attitudes to explain motivation to actions. Adding emotion would require a revised action theory. Second, if we introduce an emotion not as a natural entity but as an explanatory posit, there is a problem concerning how it will fit into Davidson's brand of rationality. Davidson admits that

... where one constellation of beliefs and desires will rationalize an action, it is always possible to find a quite different constellation that will do as well. Even a generous sample of actions threatens to leave open an unacceptably large number of alternative explanations. (1984: 160)

Davidson believes we can avoid this explosion of explanations by adhering to his decision theory brand of rationality. This gains some plausibility because in decision theory beliefs and desires are simple primitives that play irreducibly simple roles for the theory. Having one basic kind of doxastic state, and one basic kind of motivational state, seems a sparing theory — there is nothing here to shear away with Ockham's razor. But emotions are not a similarly simple primitive. They are certainly a complication to the theory, leading to actions which are inexplicable via belief and desire. But then, if we reach beyond the cut of his B-D rationality, there is no reason to add emotions into the theory instead of any other of infinitely many posits. For example, one might posit mental kind

state P which causes you to act as if some desire was not satisfied even when it is satisfied. The naturalist posits instead basic emotions because she believes basic emotions are natural, observer-independent states which can motivate actions. The interpretationist cannot hold this, and so the problem of unacceptably large numbers of alternatives rises again.

Finally, the Davidsonian interpretationist should not be allowed recourse to suppose that the agents have additional beliefs and desires to which they did not, perhaps even cannot, avow (e.g., the desire to flee as far as possible, the desire to shoot all the bullets into the dog). Such a strategy underlies Davidson's explanation of akrasia^u (1980a). This is not a strategy that Davidson has used to explain emotions; in his discussion of akrasia, Davidson wants to avoid the cases of emotional akratic actions and focus on the far more difficult cases of unemotional actions which are akratic. Still, the strategy would seem to generalize. The problem, however, is that it too would not cohere with the rest of Davidson's interpretationism. We must interpret agents in the most charitable way, trying to make them maximally consistent, and taking their own claims about their mental states as mostly right. We can posit desires that are hidden from the agent or which the agent fails to describe properly, but only by explaining how this is going to maximize coherence in the whole interpretation. We would not have this here: the emotional actions in our cases are of a type both common and commonly understood and described. There is no reason why adding new beliefs and desires would be more suitable than the explanation that refers instead to the emotion not as a judgment but as a natural, motivating state.

I can summarize by restating the problem for Davidsonian

interpretationism in a more problematic, but intuitively more compelling, way. We have already seen reason to believe that emotions (at least of the kind that concern us here) are not reducible to belief and desire, and Davidson agrees with this to the extent that he claims their logical forms are different from the logical forms of other ascribed states (1984: 156). Granting this, suppose that there was an intelligent organism which lacked emotions, and had only beliefs and desires. It would be a kind of belief-desire decision-theory thinker; popular culture examples include the character of Data.ⁱⁱⁱ Now, suppose that this thinker radically interprets the actions of humans. Using only radical interpretation, it would be unable to explain the post-functional emotional action. The best interpretation of the mental states held — and one consistent with the agent's reports, the agent's behavior in other contexts, and so on — is the one we have given and to which the agents can attest (after all, "People are in general right about the mental causes of their emotions, intentions, and actions because as interpreters we interpret them so as to make them so" (Davidson 1976: 757)), and it includes reference to emotions. (This same problem would arise for any projective version of interpretationism: if I have no emotions and I must simulate a model of the agent by ascribing my kinds of mental states, then I should fail to explain the post-functional actions.) Of course, the radical interpreter (or the pure belief-desire projector) is free to study ethology and other sciences and raise a hypothesis about emotions. Such a study would lead to the discovery of basic emotions in humans and other related animals, and also hypotheses about how these function, why they evolved, and so on. But these hypotheses now include scientific claims; we have moved into the domain of

naturalism, and posited this additional, scientifically specifiable thing: the basic emotion.

<1> The naturalist approach

I have shown that some emotions are motivational in a way that is potentially — although not always — distinguishable from the relevant intentional states that participate in and are constrained by B-D rationality. And this is in fact tantamount to common sense. Examples from folk psychology include classifying some crimes as “crimes of passion” (which I take to mean, motivated by passion). This view is not only evident in our normal discourse, but we have even institutionalized it: someone who kills in rage is treated by the law and by opinion as somewhat less culpable than someone who planned and committed a murder without anger. Here is evidenced our long tradition of treating emotions as sometimes causing actions which are somewhat independent of the control of (B-D) reason. But the leniency of the law and common opinion in these matters does not run so far as to remove all culpability; killing in rage is not like, for example, killing someone in a car accident when your brakes fail through no fault of your own. We do not treat emotions as brute, unintentional forces; in normal discourse we recognize varying degrees and kinds of voluntary control. This is consistent with these emotions being part of our mental lives, giving rise to intentional actions, but sometimes resulting in actions which fail to meet standards of B-D rationality.

I reviewed in chapters 1 and 2 some of the vast and varied scientific evidence that coheres with this view. This evidence convincingly shows that

some emotions potentially — even if not normally — are able to occur independently of the kind of cognitive skills which are necessary for B-D rationality. For Davidson and Dennett, a mental state is not the kind of thing which is directly observable, but rather is inferred from patterns of behavior. This is why both interpretationists are able to espouse versions of naturalism and still hold an irrealist belief-desire psychology. But a basic emotion is not observable only in patterns of action. The fact that an agent is experiencing a motivating, affective state is directly observable both via scientific measurements, and via expressive behaviors which are not the behavior in question. Although we may not yet have sufficient understanding of basic emotions to identify an emotional state by its autonomic or other physiological features, we can identify that an agent is experiencing some kind of affective excitation by observing such measures as galvanic skin response, blood pressure, adrenaline levels, and many other features. We are thus able at least to know, by such measurements alone, that some kind of affective state is present. I also reviewed some evidence that pancultural expressive behaviors include facial expressions (Eibl-Eibesfeldt 1973; Ekman et al 1969; Ekman and Friesen 1971; Ekman 1993) and some of these can be involuntary and operate independently of voluntary motor control (Rinn 1984). Recognizing this does not require us to interpret the action in question that results from the emotion: I can see fear expressed on Eric's face, and this is a distinct behavior from the flight which he undertakes. And this observability extends to nonhuman animals; people learn to recognize expressive features of fear in dogs, for example, and take this as a sign of a state of the dog which is not a logical construct out of behaviors but rather a potential cause of some behaviors (recalling Hebb 1946).

That we can identify emotions in nonhuman animals of many kinds is a fact which is consistent with the presupposition that our emotional capabilities evolved and as such have homologs in other animal species. Depending on how thinly one defines fear, for example, some scientists will include not only larger, more complex animals in the set of things which have the ability to fear, but even as small and relatively less complex animals as some insects (see LeDoux 1996). But many, if not all, of the nonhuman animals in question lack the kinds of capabilities, like language, that would be necessary for B-D rational thought. For Davidson, there is no fact of the matter about whether a dog has beliefs (1984: 163-64). But there can be convincing scientific evidence that a dog is in fear. The evidence surely depends upon science that began with the observation of behavior, but it extends to hypotheses about brain structures underlying fear, the evolution of fear, and substantive scientific hypothesis about the actual physiological state of being in fear. And it is these kinds of hypotheses that allow us to identify fear as a kind of capability shared by many animals. This is, in itself, not an objection to interpretationism; Dennett, who grants that the intentional stance can lead us to treat thermostats as minds, would perhaps be willing to say that flies have beliefs; and Davidson can stick to his dodge that there is no interesting fact of the matter about these issues. But this observation provides evidence that our emotional capabilities evolved in part separately from some of the relevant cognitive capabilities.

How can the affect program theory of emotion do better than the interpretationist in explaining B-D post-functional actions? Note that on the affect program theory, certain emotions (including those in our examples) can motivate an agent *to undertake a kind of activity specific to the emotion*. Eric's fear

motivates him to the action of flight; fear motivates him not just to get safely away but it also motivates the action of fleeing itself (fear, to be even more precise, *is* in part the fleeing itself). Similarly, Tim's anger motivates him to attack. He may (in some senseⁱⁱⁱ) desire to kill the dog, and he does just that, but his continued action is part of the attack. The principle characterizing feature of this realist position is that specific kinds of actions are essentially related to the emotions in question: anger is not anger that does not an attack effect or encourage, fear is not fear that does not cause flight or the preparation and motivation for flight. But it is the action, and not some particular goal, which is motivated. It may be, and probably is, right to say that some emotions evolved to satisfy some broad class of goals: it seems likely that fear probably evolved as a capability to motivate us to flee perceived threats, anger as a capability to motivate us to attack defeasible threats. But this level of explanation is distinct from that pursued in the kind of irrealist theory implicit in intentional systems theory or in various kinds of interpretationism: an emotion is not a goal-directed state in the way that a desire is.

The potential independence of basic emotions from some cognitive processes does not amount to a denial that emotions might always require or be open to something like a representational description. But the simplicity of the representational level that might be said to be required to describe flight or attack is not sufficient to integrate with the interpretationist program. A weakly representational description of emotions will in no way depend upon the notions of rationality that interpretationism utilizes as its principal explanatory constraint. The interpretationist needs basic emotions to be essentially related to

beliefs and desires to capture emotion in the web of B-D rationality.

Does it then follow that emotional actions, even post-functional ones, are arational or irrational (a distinction which can be coherently formulated once we accept that basic emotions are natural entities independent of intentional ascriptions)? If we suppose the interpretationist standards of B-D rationality, there is a sense in which we might grant that they can lead to action which is arational or irrational. Step back, however, and include in our appraisal of an agent's rationality a scientific theory of evolution and of the role of emotions, and we might see emotional actions as largely useful (as heuristics, for example) and therefore as contributing to, and indeed constituting some aspects of, our rationality (even though it would be rational not as a belief or desire or for its role in B-D rationality alone, but as a motivator to a kind of action which is, or may on balance be, under the normal circumstances rational). The hypotheses regarding the function of fear and anger given at the end of chapter 1 are examples of such functions, which themselves may be answerable to standards of rationality. I return to these issues in chapter 7. The upshot here is that, again, examining emotions in a naturalist light reveals that an irrealist view like interpretationism is too impoverished.

<1> A closing note: does interpretationism still explain content?

The primary concern of the interpretationists has been to explain belief and mental content, and not emotions *per se*. I have shown that their approach fails to explain some emotional actions and that therefore it fails to explain the relevant basic emotions; in the conclusion (chapter 13), I will further argue that

this reveals both that the presuppositions underlying the non-reductive physicalism favored by the interpretationists is a largely-misguided view of natural sciences and naturalism, and also that the interpretationists are accepting a ubiquitous fallacy concerning the importance and explanatory power of cognitive skills. However, it remains true that I have not explained, nor do I claim to have any theory of, propositional content. Could an interpretationist view still be a proper explanation of propositional content, even if not of emotional actions? I do not believe that it can. Davidsonian interpretationism, to consider what I take to be the most carefully worked-out version of interpretationism, is a tightly interconnected set of claims that includes a kind of explanation of behavior which I have here shown is, at least in part, erroneous; it also includes the claim that psychophysical laws are impossible, but this is erroneous for the same reasons. It remains to be seen what theory of content could be salvaged out of interpretationism, or other “non-reductive” views (for example, the inferential role approach of Brandom 1994), considering that we have emotional actions being explained by what is likely to be (when fully fleshed-out) a type-physicalist theory (e.g., the affect program theory), and which therefore provides a strong psychophysical relation between relational emotional actions which can play a clear role in rational action and measurable physical body states that cause them in a type-specific way. One possible approach may be that some propositional contents can indeed be type identified in some limited form in individual organisms, but that certain richer notions of content that remain significant for human beings require also that social criteria, or other criteria that cannot be given a ready, local type-physicalist explanation, be brought to bear. The issue must await much future philosophical analysis.

ⁱ Here I use the colloquial phrasing “afraid of the rabid dog,” but the reductive and doxastic cognitivists about emotions are going to be concerned with propositional forms (e.g., “Eric is afraid that the rabid dog will attack him”). My arguments will still stand if we rewrite all the sentences containing emotion terms into plausibly-corresponding propositional forms. I will return to this issue in chapter 5, however, and reveal that it is quite important.

ⁱⁱ If we assume that Karen is capable of overcoming her fear, this example is of the most obvious kind of B-D inconsistent action: akratic actions for which it is right to say that emotion provided the motivation for the action.

ⁱⁱⁱ I do allow, below, for the possibility that under some theories of desire Eric can have desires that are consequences of his emotion and which are additional to the set listed here, or even that the emotion may count as a desire. For reasons I explain — namely, that these desires require the emotion — these are not relevant exceptions to the supposed case.

^{iv} It is important to note that the normal folk psychological notion of intending is not reducible to belief and desire (see Brand 1984: 121-27), but even if it were the problem would stand: by supposition the emotional actions in question would then have the relevant kind of belief and desire as cause or constituent, and yet they are not predictable.

^v *Akrasia* is the Greek term for weakness of the will, and an akratic action is one which someone undertakes even if they know it is better for them not to do so.

^{vi} I opine, contrary to some views, that such an organism is possible. One complication is that heuristics may be practically necessary, and one effective heuristic may be emotions of the kind discussed here (we return to this theme in chapter 12). But this is not a problem for the argument, since positing such a heuristic is to posit something additional to belief and desire, which goes some way towards making our point.

^{vii} I do not believe that there is a generic motivational state corresponding to the philosopher’s notion of desire, and in fact the supposition that there is such a state may in part be reason why irrealist views like interpretationism seem plausible accounts of motivation (I discuss this in chapter 8); but since the interpretationists assume that there is such a state, or rather that the concept is an appropriate one to explain human action, I have granted their use of “desire” here and shown that even granted as is, the concept fails to account for emotional actions. But if desires are merely ascribed explanations, then under some descriptions, emotions can be counted as constituting some desire or desires; and if one supposes that there are some motivational states which are under some conditions quite like the generic notion of desire, then it is still possible that such a state would accompany an instance of a basic emotion. Thus, if one insists

upon utilizing posits of generic desires, they can still be utilized along with the notion of the basic emotions.