

Evolutionary cognitive science and emotion

The heart has its reasons of which reason knows nothing.

Blaise Pascal, *The Pensées*

All the natural minds we observe in the world around us have something in common that the classical approach completely ignores; they are all organisms, descended from a single common ancestor that lived on Earth some four billion years ago. This fact has prompted some cognitive scientists to urge their classical colleagues to pay more attention to phylogenetic questions. They recommend, in other words, that we supplement the classical focus on computation with considerations drawn from evolutionary theory. In this chapter, I examine some of these proposals for an evolutionary approach to the mind. I also argue that this evolutionary approach can help solve the problems with the classical view of emotion.

3.1. *Evolutionary cognitive science*

Classical cognitive science made little reference to evolutionary theory. The fact that all natural minds are the product of evolution was treated as purely historical matter, of little relevance to the task of understanding how minds work. In the 1980s, however, a growing band of evolutionary psychologists began to argue that the neglect of evolutionary theory had led classical cognitive science to make some important mistakes.

What are minds for?

The most serious of these mistakes, according to those working in the nascent discipline of evolutionary psychology, consisted in forgetting what minds are for. In their enthusiasm for designing machines that could prove-theorems and play chess, classical cognitive scientists had overlooked the fact that these capacities are mere by-products of the human mind. These capacities may be interesting in their own right, and designing machines with such capacities may pose fascinating technical challenges, but if our aim is to understand the fundamental properties of natural minds, theorem-proving and chess-playing are surely distractions. Human minds may be capable of such things, but they were not designed to have such capacities. Theorem and chess-playing are not the proper functions of the human mind nor of any part of it.

The difficulty of designing machines capable of solving even simple problems was enough to convince classical cognitive scientists at a very early stage that minds are very complex things. Now, according to

evolutionary theory, complex designs can only evolve by natural selection. Natural selection is not the only force driving evolutionary change, but the other forces such as random drift are not capable of generating complex functional design. Thus all natural minds must have evolved by natural selection. That is, they must have evolved because they helped organisms in certain lineages to survive and reproduce better than those in the same lineages who lacked minds. The ultimate function of all natural minds, as with any other adaptation, must therefore be to promote survival and reproduction. I will refer to this as the evolutionary theory of mind (ETM). For the sake of precision, we may sum up ETM as follows: ETM is the theory that all natural minds evolved by natural selection, and therefore that their ultimate function is to promote the survival and replication of cognitive agents.

Compatibility

It is true that classical cognitive science tended to ignore evolutionary questions about the ultimate biological function of minds, but neither did it rule them out. Classical cognitive science was simply interested in the question of how minds work – what is their design. Evolutionary psychologists, on the other hand, were interested in historical questions about how, why and when minds evolved. Unless these two questions are tied together in some way, there is no real argument to be had between evolutionary psychology and classical cognitive science. The two disciplines are asking different kinds of question. There can be no room for conflict between these research programs. They must, rather, be seen as complementary projects.

Some evolutionary psychologists do not accept this view. They argue that there is, in fact, a way of linking the synchronic question about mental structure with the diachronic question about mental evolution in such a way as to generate potential conflict between evolutionary psychology and classical cognitive science. The most famous proponents of this view are Leda Cosmides and John Tooby. In an influential paper published in 1992, these two pioneers of the evolutionary psychology movement argued that cognitive scientists could draw on *evolutionary* considerations to predict certain *design* features of the human mind and rule out others. Certain hypotheses about mental structure could be ruled out *a priori*, because they would be unlikely to evolve. Cosmides and Tooby referred to this principle as the ‘evolvability criterion’ (Tooby and Cosmides, 1992).

To demonstrate the heuristic value of the evolvability criterion, Tooby and Cosmides focused on one particular aspect of cognitive design: the question of domain-specificity. As we saw in the previous chapter, classical cognitive scientists tended to assume that the mind was a single, domain-general mechanism. In other words, it applied the same computational procedures to any and every kind of problem it encountered. Tooby and Cosmides argued that this kind of design was ruled out by the evolvability criterion; such a domain-general mechanism could not evolve, or was at least highly implausible from an evolutionary

point of view. Natural selection would, they claimed, always (or almost always) lead to the evolution of minds composed of a variety of domain-specific mechanisms.

The notion of domain-specificity needs to be spelt out in more detail, but before I do this I want to highlight the theoretical significance of the argument put forward by Cosmides and Tooby. If they are right, and the evolvability criterion does rule out domain-general mechanisms, then they will have succeeded in linking synchronic questions about mental structure with diachronic questions about mental evolution. The compatibility claim I made above about the relationship between classical cognitive science and the evolutionary approach to cognition would be refuted, and there could be genuine conflict between the two approaches. In this section, then, am not interested in the question of domain-specificity as a purely empirical matter; I am interested here only in how the evolutionary arguments for domain-specificity bear on the relationship between evolutionary psychology and classical cognitive science. In short, do the arguments about domain-specificity put forward by Cosmides and Tooby show that ETM has implications that constrain the methodology of the classical approach?

Domain specificity

Let us now return to the notion of domain-specificity. A domain-specific mechanism is one that operates only on input that meets certain conditions. Domain-specific mechanisms cannot manipulate all the representations stored in the cognitive system to which they belong. A module for vision, for example, cannot process auditory representations, and a module for face-recognition cannot process representations of plants. Modules are thus opposed to domain-general mechanisms, which can operate on any representation in the cognitive system to which they belong. Modules are 'special purpose' mechanisms, while domain-general mechanisms are 'general purpose' mechanisms

Let us call a mind composed entirely of domain-specific mechanisms '*massively domain-specific*'. Cosmides and Tooby argue that natural selection will always favour massively domain-specific minds over other kinds of mind. What other kinds of mind might there be? The obvious alternative to a massively domain-specific mind is a mind composed of a single, domain-general mechanism. As we saw in section 2.2, Jerry Fodor has proposed a third kind of mind that includes several domain-specific mechanisms and a single domain-general one (Fodor, 1983). In Fodor's model, the central executive is a domain-general mechanism that can process input from all the domain-specific mechanisms, but the latter can only process input from a single sensory source, or a single type of output from the central executive.

Evolutionary arguments for domain-specificity

Now that the terms of the debate have been clarified, let us return to the claim that natural selection tends to favour massively domain-specific minds over other kinds of mind. I will here focus on two of the main arguments that have been advanced in support of this claim:¹

- (1) Specific problems are solved more quickly by domain-specific mechanisms (Tooby and Cosmides, 1992). I will refer to this as 'the argument from specialisation'.
- (2) Domain-specific theories of mind provide the only plausible account of the evolution of the mind by showing how there could be an incremental path from very simple systems to complex minds (Marr, 1982; Brooks, 1991). I will call this 'the incremental argument'.

I will now discuss each of these arguments in more detail.

The argument from specialisation assumes that specific problems are solved more quickly by special-purpose mechanisms. This may be true, although there is not much empirical evidence for it. Even supposing it is true, however, does not licence the inference that natural selection will generally favour domain-specificity. Other things being equal, natural selection will favour a faster system over a slow one, but other things are rarely equal. Unless the environment is perfectly stable, flexibility is important too. Yet the same features that make domain-specific mechanisms fast also render them highly inflexible. Without detailed mathematical models in which the various advantages and disadvantages of domain-specificity are specified as opposing selection-pressures, we are left trading intuitions about whether or not specialisation would have been favoured during the course of human evolution. Appeals to the greater speed of domain-specific mechanisms are not convincing if they do not also take into account their decreased flexibility.

The incremental argument relies on another intuition that turns out, upon closer inspection, not be so solid. The intuition is that tightly integrated systems cannot evolve because evolution always proceeds by a series of small steps. This seems to rule out a domain-general mind, since this kind of design is so much more integrated than a massively

¹ A third argument for massive domain-specificity is also prominent in the evolutionary psychological literature. This is the argument, not that domain-general mechanisms will be out-performed by domain-specific ones, but that domain-general mechanisms would simply not have been capable of solving the adaptive problems faced by early humans in their environment of evolutionary adaptedness. Cosmides and Tooby build a persuasive argument of this type by linking it with discussions of the frame-problem. They argue that the multiplicity and variety of adaptive problems faced by our ancestors would threaten a domain-general mind with 'analysis paralysis' (not their term) because of combinatorial explosion (Cosmides and Tooby, 1992). I do not discuss this argument here since it appeals not to the evolvability criterion but to what Cosmides and Tooby call the 'solvability criterion'. That is, this argument does not turn on a putative selective advantage that favours one viable design over another, but on the putative *inviability* of one kind of design.

domain-specific architecture. It seems much easier to imagine how a massively domain-specific mind could have evolved incrementally, because we can imagine evolution proceeding by adding one domain-specific mechanism at a time. Rodney Brooks cited this advantage for domain-specific systems in connection with his own approach to robotics:

The advantage of this approach is that it gives an incremental path from very simple systems to complex autonomous intelligent systems. At each step of the way, it is only necessary to build one small piece, and interface it to an existing, working, complete intelligence.

(Brooks, 1991: 403)

David Marr also argued that a domain-specific design was more evolutionarily plausible, because domain-specific mechanisms could be 'de-bugged' individually, without rewiring the whole system (Marr, 1982). While there are important differences between the concept of domain-specificity as employed in evolutionary psychology on the one hand, and as used in robotics and AI on the other, they share the basic properties of relative computational autonomy on which the evolutionary arguments here depend.

It is not hard to see that, from the point of view of a human engineer building a robot or a digital computer, a massively domain-specific architecture is a sensible way to proceed. This is clearly what motivates the fondness for domain-specificity shown by Rodney Brooks and David Marr. But to extend this practical preference to a theoretical account of natural evolution is to make a massive leap. In particular, the incremental nature of natural selection does not bear directly on gross phenotypic features but on the individual genes, many of which are required to build a single gross phenotypic feature.

A parallel with the evolution of the body may serve to make clearer the flaw in the incremental argument. The analogy is particularly apt, since domain-specific mechanisms are often compared to physiological structures; Fodor explicitly describes them as the psychological analogue of bodily organs (Fodor, 1983). Now, nobody supposes that the incremental nature of evolution means that the human body must have evolved by adding individual organs one after the other. It would be ludicrous to suppose that an early ancestor possessed, say, just a heart and a stomach, and that this primitive species evolved by, say, acquiring first a liver, then a pancreas, and finally a brain. The steps by which evolution proceeds are much smaller, and the organism must be fully-functional at each stage of the process. Furthermore, the organs evolve in tandem with each other; it is not the case that only one organ can be modified at a time.

Just as the organ-by-organ hypothesis is implausible as an account of physiological evolution, so also the mechanism-by-mechanism hypothesis is implausible as account of mental evolution. Paul Griffiths is surely right

when he states that it is 'implausible that our brains evolved by adding separate mechanisms subserving new functions' (Griffiths, 1999: 51). The mind may be massively domain-specific, but if it is, the various mechanisms surely evolved in parallel, just like the organs of the body. Conversely, there is no reason why a domain-general mind could not have evolved incrementally by a process of gradual expansion. Thus the incremental nature of natural selection does not predict a massively domain-specific mind.

I have argued that the two evolutionary arguments for domain-specificity do not work. The claim that natural selection tends to favour domain-specific minds, therefore is not proven. I conclude that Cosmides and Tooby have not succeeded in linking the *structural* question of how natural minds are designed to the *evolutionary* question of how they evolved in such a way as to generate potential conflict between the evolutionary and the classical approaches to cognition. The two research programs are perfectly compatible.

Evolutionary psychology and artificial intelligence

Evolutionary psychology is something of an odd-man-out among the various non-classical approaches that I discuss in this thesis. In line with the design-based approach of classical cognitive science, situated cognitive science and dynamical cognitive science both involve intimate links between theory (theory of mental structure) and practice (the practice of building artificial minds). Most evolutionary psychologists, on the other hand, have been exclusively concerned with theories of human mental structure, and few have attempted to translate their models into working machines.²

This might seem to exclude evolutionary psychology from cognitive science, at least if we go by the definition of cognitive science that I proposed in chapter one. There, I defined cognitive science as any approach to the study of the mind that (1) accepts the computational theory of mind, and (2) adopts a design-based approach. Evolutionary psychology certainly satisfies the first of these conditions; the discipline owes its very name to the desire for a label that would both mark the rejection of the rather behaviouristic approach typical of much sociobiology, and signal the adoption of an explicitly computational approach (Cosmides and Tooby, 1987; Caporael, 1989). However, it is not clear whether evolutionary psychology meets the second condition. With a few notable exceptions, most of those who call themselves evolutionary psychologists have not, as yet, been involved in designing artificial minds. They would therefore seem to lie outside the field of cognitive science, at least as I have defined it.

However, this conclusion is too quick. My definition of cognitive science, it will be recalled, does not specify that all cognitive scientists must take an

² Notable exceptions include Geoffrey Miller, Gerd Gigerenzer and Douglas Kenrick.

active part in *building* artificial minds. It simply states that cognitive scientists must adopt a design-based approach. As I noted in the introduction, this condition is fulfilled whenever researchers propose models of the mind that are computational enough to permit computer programs to be *readily* designed on the basis of the models. Many of the domain-specific mechanisms proposed by evolutionary psychologists take such a form; they are not specified in terms of any programming language, but they are often spelled out in a form that would be relatively easy to convert into a computer program. The models proposed by evolutionary psychologists thus count, on my definition, as fully cognitive, and evolutionary psychology is firmly within the fold of cognitive science.

Even so, it seems a shame that evolutionary psychologists have not taken more interest in translating their models into real machines. The tools of artificial intelligence and computational modelling might well offer them ways of testing their hypotheses and thus enable them to answer the common charge of telling 'just-so stories'. Critics of evolutionary psychology frequently dismiss it on the grounds that it promulgates untestable theories. Evolutionary psychologists acknowledge that there are methodological difficulties posed by investigating the history of the mind, but point out that most of these difficulties are not particular to their discipline. Most of them are common problems faced by all those who do wish to investigate evolutionary hypotheses, so to be consistent the critics should also dismiss the whole of evolutionary biology. Such general defences, however, would be strengthened if evolutionary psychologists could point to *experimental* ways of testing their hypotheses. Artificial intelligence could supply evolutionary psychology with just such experimental techniques. There are, in particular, some relatively new techniques that would be particularly relevant, because they explicitly address evolutionary questions.

Artificial life and evolutionary robotics

One of these new techniques is known as artificial life (or simply as 'A-Life'), a name coined by Christopher Langton in 1986 (Langton, 1986). Instead of trying to build complex machines in the normal way, by forward planning, researchers in A-Life attempt to model the process of natural selection. They remove the human engineer from the process as much as possible by using a random process to generate various alternative designs, and then allow the better designs to replicate. Errors are deliberately built into the replication process to mimic the mutations that occur naturally when biological systems reproduce. Repeated rounds of differential replication lead to increasingly refined designs, just like natural selection.

The advantage of this method is that it can lead to extremely novel designs. The random nature of the 'mutations' means that A-life is unhindered by the assumptions and prejudices that human engineers bring to any task, and so can explore regions of design space that humans might never find on their own. Some even argue that the complexity of

some design problems is such that foresight is practically impossible, so that selection is in fact the optimal search strategy for exploring the hyperspace of all possible designs.

An example of this evolutionary approach to engineering was recently provided by Pablo Funes and Jordan Pollack of Brandeis University. Funes and Pollack constructed a program which generated random LEGO designs for various structures such as a two-metre bridge and a table capable of supporting a one-kilogram weight. The program also analysed these designs by using various algorithms for measuring torque, stress, leverage, etc. The bad designs were then eliminated, and the remaining ones fed back into the process, where they were modified by further random 'mutations', analysed again, and so on. Using this completely unsupervised process, the program was able to produce a sophisticated bridge with a cantilevered design in a day and a half. Funes and Pollack tested the various structures designed by their program with actual LEGO bricks, and found that they were all structurally sound.

A-Life is an umbrella term that embraces many other projects besides designing structures like bridges. A-Life methods have been used, for example, to model RNA replication and population dynamics, and even to produce works of art. Here, however, I shall concentrate on just one branch of A-Life research – namely, that which is concerned with the design of artificial autonomous agents, or 'animats'. The class of animats includes both mobile robots that inhabit the real world and simulated agents embedded in virtual environments.

One strategy for designing animats is to use the same kind of evolutionary approach as that used by Funes and Pollack. In one classic example of this approach, Thomas Ray designed a virtual world called *Tierra* and populated it with a simple digital organism. This was a simple self-replicating program (or 'genetic algorithm') which occasionally made mistakes in the copying process. These 'mutations' led to an increasingly diverse population of digital organisms. Competition for limited memory space ensured that there was differential survival. The conditions for natural selection were therefore all in place, and Ray was able to observe numerous cases of digital evolution complete with virtual viruses, parasite resistance and other surprisingly 'natural' features (Ray, 1992).

Tierra is only a virtual world, and thus subject to the criticisms of roboticists like Rodney Brooks, who argue that it is all too easy in such simulations to make some crucial but unnoticed simplification that renders the simulation invalid (see chapter three). In order to avoid this potential danger, Brooks recommends that cognitive scientists work with artificial autonomous agents that inhabit the real world – i.e., with robots. When this recommendation is combined with the A-Life approach to designing such agents, the result is a strategy known as 'evolutionary robotics' (Wheeler, 1996). In evolutionary robotics, genetic algorithms are usually used to develop robot control systems ('robot minds'), but there is no reason why they should not also be used to design better robot bodies. This is, in fact,

the way Funes and Pollack envisage their engineering program being used in the future. If a computer can design sound bridges and tables, then it is only a short step before it can design working computers and robots. It is another short step from this to the idea that robots will actually build the robots they design. Artificial autonomous agents that make copies of themselves need not be confined to virtual worlds. They could also come to exist in the real world, giving rise to a genuine lineage of robots evolving by natural selection.

Such a scenario is currently beyond our technical capabilities. However, the basic principle of using natural selection to design autonomous agents is sound, and has at least been tested in virtual environments such as Ray's *Tierra*. The hope of researchers in Artificial Life and evolutionary robotics is that these methods may one day give rise to an artificial agent with humanlike intelligence. Perhaps these methods can provide a way, then, for evolutionary psychologists to test their hypotheses about mental evolution. They can simply watch it happen *in silico* and observe what kinds of mind tend to evolve.

Evolution and contingency

One problem with using these evolutionary approaches to design artificial minds is that this might tell us nothing about the design of real minds. Evolution is a notoriously contingent process. The important role played by historical accident in evolution has led Stephen Jay Gould to remark that if we could rewind the tape of biological history and start it again, the outcome would probably be very different. Not only might there not be humans, Gould suggests; there might not even be anything like mammals. Even if we did succeed in creating an intelligent machine by means of some evolutionary design process, therefore, its mental structure might be very different from our own. If the point of building an artificial mind is to increase our understanding of real minds, using an evolutionary design-process to build one might be a dead end.

On the other hand, evolution may not be quite so contingent as Gould suggests. If we could rewind the tape of biological history and start it again, perhaps we *would* find similar kinds of outcomes. Since life on earth has only evolved once, it seems that there is no way of arbitrating between these different possibilities; we are left trading intuitions. However, researchers in A-Life dispute this. They claim that we *can* rewind the tape of biological history and re-run it thousands or even millions of times.

By running programs like *Tierra* over and over again, perhaps varying the initial parameters occasionally, we might just be able to discern various *constants* in evolution. Computer simulations of evolution might provide a way of testing Gould's claim about the radical contingency of evolution. What counts as similarity and difference depends, of course, on your frame of reference. If we are concerned with details, such as the number of digits on a limb, then perhaps we *will* find a different outcome each time

we run our computer simulation of evolution. However, if we use a less fine-grained taxonomy, we may find the same broad classes of organism turning up every time we let our virtual world evolve. This line of thought is what prompted Ray to note that he found virtual viruses evolving in *Tierra*. These viruses did not use RNA, and were not encased in a protein shell; they were simply strings of digits on the computer's hard disk. However, they had certain important properties in common with natural viruses. They could not, for example, replicate in isolated culture, but only when cultured with normal (self-replicating) creatures. Like natural viruses, the artificial parasites executed some parts of the code of their hosts. As in the real world, some potential hosts in *Tierra* evolved immunity to the virtual viruses, and some of the viruses then evolved mechanisms to circumvent this immunity (Ray, 1992: 124).

Ray's analysis of evolution in *Tierra* supports the idea that, while the details may change, the underlying patterns may be the same whenever evolution occurs. Given enough time, we may find that every evolutionary process tends to produce the same basic classes of organism, filling the same kinds of niche. If we make our taxonomy coarse enough, this statement will become trivially true. If we use very abstract ecological categories, such as parasite and host or predator and prey, for example, re-running the tape of evolution will almost certainly produce similar outcomes. Thus the claim that evolution always produces parasites may not be very interesting since the term parasite is so broadly defined. Conversely, if we find that evolution only rarely produces animals with five digits on each limb, this may not be very interesting either, because this kind of detail is of no particular consequence. The interesting questions focus on categories that are neither too general nor too specific.

I suggest that it is just these kind of interesting questions, at the right grain of analysis to make them worth investigating by computer simulations of evolution, that are posed by much of the work in evolutionary psychology. The claim that massively domain-specific minds tend to be favoured by natural selection has not been demonstrated on purely theoretical grounds, but it might be demonstrated on experimental ones. If we ran hundreds of simulations of human evolution, or animal evolution in general, now and again varying the initial parameters, and found that domain-specific minds were almost always the end product, this would provide strong evidence in support for the idea that natural selection tends to favour such minds. Such a finding would therefore provide encouragement for cognitive scientists to model the human mind by designing domain-specific machines. Conversely, if we found that domain-specific minds were only rarely produced by computer simulations of cognitive evolution, this would provide grounds for betting that the mind was better modelled by domain-general machines. Thus, while ETM is compatible with CTM, evolutionary approaches to cognition are not redundant; in particular, they can suggest interesting ways of narrowing the search space for the design that best approximates that of the human mind.

Evolutionary psychology or evolutionary cognitive science?

This suggestion would provide a way of bridging the gap that currently separates evolutionary psychology from artificial intelligence. I have already noted that evolutionary psychologists have not yet engaged in much explicit dialogue with researchers in artificial life and evolutionary robotics. This is a shame, because, as I have tried to show in this section, they are natural partners. Evolutionary psychology can be seen as investigating the evolution of *natural* minds, while A-Life (or at least some branches of A-Life) and evolutionary robotics can be seen as investigating the evolution of *artificial* minds. If we want to adopt the methodological maxim of cognitive science, according to which we build artificial minds in order to understand natural ones, evolutionary psychology and A-Life should be seen as two sides of the same coin. That is why I propose the umbrella term, 'evolutionary cognitive science', to cover the potentially fruitful intersection between evolutionary psychology, on the one hand, and A-Life and evolutionary robotics on the other. In the next section, I argue that evolutionary cognitive science would have many important things to say about the emotions.

3.2. Evolutionary approaches to emotion

In section 1.2 we saw that cognitive science faces a dilemma: it must either reject the cognition/emotion distinction, or find an alternative way of explicating this distinction to that proposed by Hume. The classical models of emotion developed by the appraisal theorists did not solve this dilemma. They did succeed in providing a representational account of emotion, by construing emotions as kinds of belief or judgement, but they did not specify any independent criterion for distinguishing emotional types of judgement from other non-emotional (cognitive) types.

As Zajonc said (or meant to say), if cognitive science is to solve the dilemma, it needs further conceptual resources. In this section, I argue that the evolutionary approach can provide these resources. Evolutionary theory can help cognitive science to deal with emotion in the same way that it helps cognitive science to deal with other mental processes: by bringing in the question of *function*. Before discussing the functions of the emotions, however, I will argue that the evolutionary approach can solve another problem posed by the classical account of emotion – the problem of what emotions are *about*.

What are emotions about?

Hume, it will be recalled, argued that emotions are non-representational. Unlike thoughts or beliefs, they are not *about* anything; they just *are*. So, for Hume, it would make no sense to ask whether or not an emotion was 'correct' or 'true'.

Classical cognitive scientists interested in emotion rejected this Humean view and replaced it with a representational account of emotion. Emotions, on this account, are particular kinds of judgement or belief. Psychologists working with appraisal theory and philosophers working with the propositional attitude theory both went about showing how particular emotions could be construed in this way. If emotions are representational, however, it follows that there must be some way of assessing their truth value. And yet to many people this seems distinctly odd. Even if we disagree with Hume's view of emotion, we can still feel the tug of his intuition when he writes that it is 'impossible ... that this passion be oppos'd by, or be contradictory to truth, or reason' (Hume, 1734: 415). Most people, after all, would feel rather puzzled, if not downright offended, if you asked them whether, say, their embarrassment was *correct* or not. If they bothered to reply, they would probably say that it *must* be correct, because they feel it. Such a reply implicitly endorses Hume's view of emotion, and challenges the classical cognitive scientist to point to some *external* system by which the truth of an emotion can be judged.

The classical cognitive scientist would be at a loss here. If one adopts an explicitly evolutionary approach to the study of the mind, however, there is a possible solution; the environment can serve as the external system. In other words, a particular instance of an emotion in a particular organism can be judged 'correct' when it fulfils the proper function of that type of emotion. If the function of fear is to help the organism avoid danger, for example, then an organism is right to feel fear on this occasion if and only if there is an immediate danger. Otherwise, this instance of fear will be 'incorrect'.

Functional accounts of emotion

Let us return now to the question of what functions are served by particular emotions. In order to understand this question, it is first necessary to be clear about how functional statements in biology should be understood. The meaning of functional ascriptions has generated substantial philosophical interest during the past few decades, and a rich literature has grown up around the topic. An extensive analysis of this literature would take me too far away from the subject of this thesis, so I will limit myself to summarising the main conclusions that have emerged from it.

There is now a good deal of consensus among philosophers of biology that functional statements in the biological sciences are usually to be understood as making a historical claim about causes.³ More specifically,

³ Functional statements in psychology may be treated quite differently. In particular, when functions are ascribed to mental processes, as in functional decomposition (see chapter one), this does not imply any historical or evolutionary claim. Rather, statements about cognitive functions are analysable in purely synchronic terms, as reducible to statements about the causal role that cognitive processes play in the overall mental economy. Some philosophers, such as Ruth Millikan, have attempted to ground these statements about cognitive functions in statements about biological functions, but this 'teleo-semantic' approach is an independent issue.

a statement to the effect that ‘the function of trait x is y’ is to be understood as shorthand for the claim that ‘the reason this population has trait x is because x helped its ancestors to survive and/or reproduce by doing y’. For example, to say that ‘the function of the heart is to pump blood’ is to say that the reason we have hearts today is that hearts helped our ancestors to survive by pumping blood around their bodies. Functional statements thus imply that the trait in question conferred a selective advantage on those ancestral organisms who had the trait, and that this selective advantage played a causal role in the proliferation of the trait among the daughter population.

In proposing functional accounts of human emotions, therefore, evolutionary psychologists are not claiming that emotions still help contemporary humans to survive and/or reproduce, but simply that they helped our recent ancestors to do so. Whether or not emotions are still selectively advantageous is left open by the functional hypotheses advanced by evolutionary psychologists.

There is a high degree of agreement between evolutionary psychologists about the functions of many emotions. I have based the following brief summary mainly on the work of Leda Cosmides and John Tooby (Tooby and Cosmides, 1990), but similar evolutionary accounts of emotion have been put forward by Paul Ekman, Robert Plutchik, R. S. Lazarus and Randolph Nesse (Plutchik, 1980; Nesse, 1990; Lazarus, 1991; Ekman, 1992). All these researchers argue, uncontroversially, that fear helps animals to survive by avoid predators by fleeing (to escape) or freezing (to avoid being spotted). Most agree that disgust is clearly of value in helping animals to avoid ingesting or touching substances that may be poisonous or infectious. There is similar agreement that surprise alerts animals to a change in the environment, while anger readies them for combat.

If we are to provide a functional definition of emotion, then we must abstract away from the functions of particular emotions and ask what they all have in common. The proposal I will argue for here is that all emotions are able to achieve their particular functions only because they fall under a more general functional category: they are all interruption mechanisms. Fear makes you stop what you are doing *when you detect danger*. Anger interrupts ongoing activity in order to deal with possible combat. Disgust stops you eating potentially dangerous food or going near sources of infection. And so on.

Interruption mechanisms

The idea that emotions are interruption mechanisms can be traced back to Herbert Simon, who proposed a functional definition of emotion in a short but fascinating paper in the 1960s (Simon, 1967). Simon started from the simple observation that there is a limit to the amount of things that any agent can do at any one time, whether it be an animal or a robot. Therefore, if the agent has more than one goal, it must divide its time up wisely, allotting the right amount to each activity in pursuit of each of each

goal. However, unless the environment is completely stable and benign, the agent must also remain alert to external changes that may require a rapid change of activity. Suppose, for example, that a robot has the following two goals: *first* to collect rock samples from an asteroid and analyse them *in situ*, and *second*, to bring these samples safely back to earth. Now imagine that such a robot is sitting happily on the asteroid, conducting some chemical test on the rock it has just picked up, when suddenly a piece of debris comes hurtling towards it. Unless the robot has some kind of 'interruption mechanism', it may succeed in its first goal, but fail dismally in the second.

Simon proposed that emotions were just such interruption mechanisms. He meant this as a definition. In other words, the word 'emotion' is the name we have given to these interruption mechanisms when we have observed them in ourselves and other animals. According to Simon's functional definition, emotions are those mental processes that generally work to interrupt activity in rapid response to a sudden environmental change. More recently, Keith Oatley has developed Simon's ideas into a theory of emotion according to which 'an emotion is a psychological state or process that functions in the management of goals ... it is an urgency, or prioritization, of some goals and plans rather than others ... (that) can interrupt ongoing action (Oatley, 1999: 273; see also Oatley, 1996). Oatley's theory drops Simon's emphasis on reaction to sudden external changes, and thus allows that there might be interruption mechanisms that are triggered by less sudden changes and by internal events.

Objections to the interruption theory

Let us call the idea that emotions can be functionally defined as interruption mechanisms 'the interruption theory'. It is now time to examine some possible objections:

- (i) The interruption theory is too broad; it would class every mental process as emotional.

If any mental process can interrupt any other, then we will not have succeeded in finding an alternative, non-Humean way of distinguishing cognition and emotion. If we are to pick out emotions as a distinct class of mental processes on the basis of their capacity to interrupt other mental processes, then there must be some mental processes that are *incapable* of interrupting others (and, presumably, we will have to identify these with the class of 'cognitive' processes).

So, are there, in fact, any mental processes that never interrupt others? It is certainly possible to imagine how this might be the case. If mental processes could be organised in a simple hierarchy, such that a process at one level in the hierarchy could only interrupt those in the level above it, then those at the top of the hierarchy would clearly be incapable of interrupting any other. This is, in fact,

the kind of mental architecture proposed by Simon in his paper on emotions. It is also the basis of many architectures in contemporary behaviour-based robotics, which often consist of many autonomous layers. In these systems, the robots' behaviour is only controlled by one layer at any one time. The highest layer is the default control layer. In other words, so long as none of the lower layers is triggered into action, the highest layer directs the movement of the robot. However, while the robot is under the control of one layer, all the lower layers are still alert to possible stimuli. If the relevant stimulus triggers one of the lower layers into action, it automatically takes over.

Such hierarchical architectures are not restricted to robots, however. Jaap Swagerman has implemented such an architecture in a program he wrote for a desktop computer. The program is called ACRES (an acronym for 'Artificial Concern REalisation System') (Swagerman, 1987). ACRES is a database containing information about emotions and the situations that give rise to them, but these data are not the relevant point here. More important than the information contained in the database is the fact that ACRES is also a very sophisticated interface. ACRES has multiple concerns, and now and again it examines these concerns to see if any of them requires action. The concerns are, in decreasing order of importance: to stay 'alive' (switched on), to get fast input, to get accurate input, to get varied input, and so on, down to servicing database queries and turning on and off its debugging procedures. While the user is interacting with ACRES (asking it for data, inputting data, ending the session, etc.), the program scans its list of concerns, beginning with the most important, and takes appropriate action where necessary. For example, if it has not learned anything new for a while, its concern for getting varied input will trigger a request for the user to tell it something new. If the user does not comply with the system's wishes, he is gradually given the cold shoulder, first being refused permission to change the database, and eventually being denied access to ACRES altogether.

Users interacting with ACRES report that the 'emotional behaviour' of the program feels quite realistic (Moffat, Frijda et al., 1994). On the interruption theory, this impression is understandable; the program really *does* have emotions. Each of the multiple concerns of ACRES can potentially interrupt ongoing activity. Each concern is assigned a fixed importance index, and these are ranked in a simple hierarchical fashion so that, although goals can conflict, there is always a simple algorithm for determining which takes precedence.

If emotions really are interruption mechanisms, then all the layers in this kind of architecture instantiate emotions, and only the highest layer could be said to be truly 'cognitive' (in the sense of being unemotional). *Prima facie*, then, this 'default view' of cognition

seems like a fairly good approximation of what we mean when we speak of pure thought, or pure reasoning, in humans. Only when we are *not* prey to a particular emotion do we usually attribute such purely 'cognitive' processes to ourselves.

Every layer must have some kind of goal, though. So, on the interrupt theory, emotions cannot simply be equated with goals or drives. The goal of the top layer must be something that never interrupts other goals. Curiosity or 'interest' meets this criterion. This is why Izard (1979) regards it as the default state of the organism. Izard calls interest an emotion, but on the interruption theory it would not be classed as such. It would, in fact, be the defining feature of unemotional (i.e. cognitive) processes!

- (ii) The interruption theory is too broad; it may not class every mental process as an emotion, but it still allows us to call some things emotions when they are not.

What about things like hunger and pain? These are not usually described as emotions, but they can clearly interrupt other mental processes. This objection can be met in a number of ways. One way would be to refine the interruption theory by adding an extra clause to our definition of emotion, so they are defined not simply as interruption mechanisms, but as interruption mechanisms *of a certain sort*. We would then have to find a way of distinguishing between interruption mechanisms such as hunger and pain, on the one hand, and interruption mechanisms of a more obviously emotional kind on the other. I find this response unattractive, however, as I can think of no principled way to make such a distinction. I therefore prefer to adopt an alternative response to objection (ii).

The alternative response I prefer is to say that we were wrong to exclude things like pain and hunger from the class of emotions. It is not uncommon for us to revise our pretheoretic use of terms in the light of later scientific theories. If the range of things we refer to as 'emotions' in everyday, pretheoretic usage overlaps to a large extent with the class of interruption mechanisms, it is reasonable to see the pretheoretic term as an initial approximation to the scientific theory. We can then accept that the pretheoretic use of the term wrongly excluded (or include) a few things that we now think share the defining properties identified by our the scientific theory. In other words, so long as most of the things we designate as 'emotions' in our pretheoretic way are interruption mechanisms, then there are good grounds for arguing that our pretheoretic judgements about the unemotional nature of pain and hunger are just plain wrong.

The same response can be used to tackle cases of non-human emotion. Many people seem quite happy to attribute emotions to other primates, and indeed to many other mammals, but they become

less happy to attribute emotions to other species less related to us. Most people would probably baulk at the idea that *worms* have emotions, for example. Yet worms clearly have interruption mechanisms, so, on the interruption theory, they can truly be said to have emotions. As with the case of hunger and pain in humans, we could deal with this objection by adding an extra clause to our definition of emotions as interruption mechanisms. For example, we could say that interruption mechanisms only deserve to be called emotions when they are found in creatures above a certain threshold of cognitive complexity. This is too arbitrary, however. The fact that such arbitrary post hoc modifications are required to bring the interruption theory into line with the common usage of the term emotion simply indicates that the pretheoretic use of the term is too anthropocentric, too obsessed with the local features of the various interruption mechanisms we find in ourselves. We should therefore not let the common reticence about attributing emotions to lowly creatures like worms stand in the way of a scientific account of emotion. So long as the interruption theory has hit on the element that we were groping towards in our pretheoretic use of the term, then we should not hesitate to amend our usage when it conflicts with the interruption theory.

I think that it does hit on such an element. The interruption theory provides a precise way of explicating the notion that lies at the heart of our pretheoretic notion of emotion. This is the idea that emotions can take us over against our conscious volition. This idea is implicit in the older term 'passion', which comes from the same root as 'passive'. We are, in a very real sense, often passive 'victims' of our emotional reactions.

(iii) The interruption theory does not allow for top-down influences of cognition on emotion.

Humans are not entirely at the mercy of their emotions. Emotions often have an imperious, automatic quality to them, but no so much as to render them always impervious to voluntary control. Yet the interruption theory does not seem to allow for such top-down influences. This objection could be met by building some kind of variable threshold into our model. If lower levels could only interrupt higher levels when their signal exceeded some given threshold, and if higher levels could exert some influence on the level of this threshold, then, there would be some measure of top-down influence without thereby abolishing the distinction between mechanisms that are capable of interruption and those that are not.

(iv) The interruption theory applies only to a subset of those things we call emotion.

All the emotions discussed so far in this chapter are among the so-called 'basic emotions' identified by Paul Ekman. These include fear,

anger, surprise, disgust, joy (or happiness), and distress (or sadness). According to Ekman, basic emotions are typically automatic, reflex-like responses of rapid onset and brief duration (Ekman, 1992). They thus seem well suited to being described as interruption mechanisms.

But not all emotions are 'basic' in this sense. Love, guilt, shame, jealousy and sympathy are certainly emotions, but they do not possess same suite of features that, according to Ekman, define *basic* emotions. For this reason, Paul Griffiths has argued that they deserve to be treated in a class of their own, which he refers to as 'the higher cognitive emotions' (Griffiths, 1997). Perhaps these emotions are not interruption mechanisms. If so, then we must either seek some other definition of emotion, or accept that emotion is not a natural kind. Griffiths prefers the latter option, but I think he gives up too quickly on the project of finding a good definition of emotion.

The first thing to note, in response to this objection, is that it may be misleading to lump all non-basic emotions together into a single class.⁴ Just because an emotion does *not possess all* the characteristics that define basic emotions does not mean that it *lacks them all*. The properties that Griffiths attributes to higher cognitive emotions are the all contraries of those that Ekman attributes to basic emotions. Thus, according to Griffiths, all higher cognitive emotions take longer to build up and die away than basic emotions, all involve much more cortical processing than basic emotions, and all lack universally-recognisable distinctive facial expressions. There are reasons to doubt that these features correlate as highly as those that define basic emotions, but I will not go into them here. In order to avoid the dubious assumption that all non-basic emotions form a natural kind, as robust as the natural kind formed by basic emotions, I will not refer to them as 'higher cognitive emotions', but simply as 'non-basic emotions'.

Secondly, it is worth remembering that Ekman proposed the distinction between basic and non-basic emotions, not because he thought that emotions could be divided into two robust natural kinds, but because he was trying to convince his fellow anthropologists that some emotions, at least, were universal and innate. By picking out a number of properties shared universally by some emotions, he was able to mount a persuasive argument against the cultural theory of emotion, which viewed all emotions as learned phenomena and thus culturally-specific. The historical context of

⁴ Griffiths does not say quite this. He allows that there may be a third class of emotions that are culturally-specific. However, he still thinks that the *pancultural* emotions can be divided into two classes, basic emotions and higher cognitive emotions, and implies that, for every dimension on which basic emotions and higher cognitive emotions can be compared (other than the pancultural/culturally-specific dimension), they take opposite values.

the concept of basic emotions should warn us against granting too much metaphysical weight to the basic/non-basic distinction. In particular, we should not assume that, just because the interruption theory of emotion fits the basic emotions, it therefore does *not* apply to non-basic emotions. The question of whether non-basic emotions can also be treated as interruption mechanisms, and therefore the question of whether the interruption account amounts to a general theory of emotion, cannot be settled simply by appealing to the fact that a subset of emotions share the properties identified by Ekman.

To settle the question of whether or not emotions like love and guilt can be described as interruption mechanisms, we need to ask what their particular functions are, and then ask whether or not these functions are plausibly regarded as species of the more general functional category of interruption mechanisms. This is not so easy; functional accounts of the non-basic emotions have been much thinner on the ground than functional accounts of basic emotions. The only person to put forward an extensive theory of why (at least some of) the non-basic emotions evolved is the Cornell economist, Robert Frank. In his book, *Passions within Reason*, Frank argues that many non-basic emotions evolved to help our recent ancestors solve various kinds of 'commitment problem' (Frank, 1988).

Commitment problems arise whenever an agent needs to make a credible threat or promise. Threats and promises are vital to successful co-operation, but threats can be empty and promises defaulted upon, so the agent who wishes to co-operate must convince others that he is sincere. One way for him to do this, argues Frank, is to provide evidence that he is committed to carrying out the threat or promise willy-nilly, even if it becomes disadvantageous for him to do so. He needs, in other words, to show that he is 'handcuffed' to carrying out the promise or threat. This may be termed 'the handcuff principle'. According to Frank, many non-basic emotions provide both the handcuff itself (in the form of an uncontrollable feeling) and the evidence that such a handcuff is in place (in the form of physiological signals, such as sweating or blushing).

Take guilt, for example. It might seem that feeling bad when you cheat is not very advantageous in a world governed by the iron law of the survival of the fittest. Yet if others know that you feel bad about cheating, they will be more likely to co-operate with you in joint-ventures that require trust. The fact that the feeling of guilt cannot be easily swayed by a calm, rational assessment of self-interest is vital. There are many occasions in life when it is possible to take a benefit without paying the corresponding price, and without being detected. In such a situation, the most rational thing to do (as defined by rational decision theory) is to cheat. However, when one takes a broader view, the calculation of costs and benefits changes

somewhat. The feeling of guilt can force one to take this broader view, when reason might otherwise focus on the short-term.

Frank illustrates his analysis of guilt with the following example. Consider two people, Smith and Jones, who wish to engage in a potentially profitable joint-venture, such as starting a restaurant. Their potential for gain arises from the advantages associated with the division of labour. If Smith is a talented cook, and Jones is a shrewd manager, they can use their respective skills to launch a successful joint venture that pays each of them more than they would gain from working alone. The problem is that each will have opportunities to cheat without being detected. Smith can take kickbacks from food suppliers, while Jones can fiddle the accounts. If only one of the partners cheats, he does very well, while the other does poorly. Thus self-interest dictates that cheating is the best policy, and, if both are rational agents, both end up cheating. With both parties cheating, however, each does worse than they would do if both were honest. This is simply a version of the famous 'prisoner's dilemma' so beloved by game theorists (Axelrod, 1984). If Smith and Jones could make a binding commitment not to cheat, both would profit by doing so. The problem thus reduces to that of how to make a credible commitment (Frank, 1988: 4-5).

Frank proposes that the emotion of guilt is one way of solving this commitment problem. If a person feels guilty whenever he cheats, this can cause him to behave honestly even when he knows that he could get away with cheating. And if others know that he is this type of person, they will seek him out as a partner in joint ventures that require trust. This depends, of course, on there being reliable cues that indicate the presence of guilt. Only if there is some signal that is good evidence for a conscience, such as blushing when one feels guilty, will people know that you are trustworthy. These signals must be hard to fake, otherwise they would not be reliable. Frank argues that natural selection has built such hard-to-fake signals into human physiology precisely to solve the commitment problem.

The irony of the prisoner's dilemma is that the failure to pursue self-interest actually leads to genuine advantages, at least when self-interest looks only to the short-term. On Frank's account, emotions like guilt save us from the pitfalls of using reason alone. However, just because such emotions work against the dictates of human reasoning does not mean that they are 'irrational', in the technical sense of flouting the principles of rationality theory. When considered in the context of a one-shot game, they are clearly irrational, but when set in the context of a series of repeated interactions with the same players they exhibit a kind of 'global rationality' that saves human *reasoning* (not pure *reason*) from itself.

Frank's analysis of guilt as a solution to a commitment problem can be extended to other non-basic emotions such as the 'sense of fairness', vengefulness, and romantic love. Forming a stable pair-bond for the purposes of rearing children is another example of the commitment problem. Jack and Jill may consider each other as a suitable mate, but forming a stable pair-bond requires a substantial investment of time and resources, and each fears that that this investment could be undercut if the other were to leave for an even more attractive partner in the future. Without reasonable assurance that this will not happen, neither will be willing to make the investments required for a successful pair-bond (Frank, 1988: 49). The emotion of romantic love is a solution to this problem. If Jack commits himself to Jill because of an emotion he did not 'decide' to have (and so cannot decide *not* to have), an emotion that is reliably indicated by tachycardia and insomnia, then Jill will be more likely to believe he will not default on his commitment than if he had chosen her after coolly weighing up her good and bad points. 'People who are sensible about love are incapable of it,' wrote Douglas Yates (Pinker, 1997: 418).

Emotions like guilt and romantic love are termed 'higher cognitive emotions' by Griffiths because they show much more sophisticated processing than basic emotions, and because they are much more integrated with other cognitive processes such as those leading to long-term planned action – not because of any susceptibility to conscious control. Lack of conscious control is, in fact, vital to these emotions, since an ability to control them by rational deliberation would defeat their purpose. If the function of these emotions is indeed to 'save rationality from its own pitfalls', as Frank argues, they *must* be fairly autonomous. Both the feeling and the signal must be hard to fake. In other words, these emotions too must be capable of *interrupting* thought when they detect some reason for doing so. There are good grounds, then, for thinking of higher cognitive emotions as interruption mechanisms too.

Emotions, domain-specificity, and modularity

I do not pretend to have dealt with all possible objections to the interruption theory, but I hope I have at least dealt with some of the most obvious ones. I now wish to return to the issue of domain specificity in order to clear up some terminological confusion that has at times side-tracked those interested in this question.

The interruption theory implies that emotions must be subserved by domain-specific mechanisms. Only when a mind is composed of several distinct subsystems, each attending to a different kind of input – only, that is, when a mind was massively domain-specific – could it have interruption mechanisms. If the interruption theory is correct, then, a mind could not have emotions unless it was massively domain-specific (though the interruption theory does not rule out the possibility that such a mind could

have a domain-general mechanism too, subserving cognitive processes). This argument is very similar that put forward by Cosmides and Tooby in their 1990 paper on emotions (Tooby and Cosmides, 1990).

In fact, there are good reasons for thinking that the mechanisms subserving emotions must not only be domain-specific, but that they must also have most, if not all, of the other properties that Fodor takes to define 'modules'. I hesitate to use this term, as it is used in such a variety ways by various sections of the cognitive science community that it tends to impede communication rather than to facilitate it. I will therefore spell out exactly what I mean by it before I go on.

However other people may use the term, I follow Fodor in thinking of a module as a computational mechanism (computational in the sense defined in chapter one, i.e. as something that performs transformations on representations) with the following nine properties (c.f. Fodor, 1983):⁵

- (1) *Domain specificity* – modules only perform transformations on representations that fall within a certain domain.
- (2) *Mandatory operation* – modules are automatic, like reflexes.
- (3) *Inaccessibility to conscious introspection* – the intermediate transformations performed by modules are not accessible to consciousness.
- (4) *High speed* – modular mechanisms are much faster than non-modular ones.
- (5) *Informational encapsulation* – the database and program of a module is not available for use by other mental mechanisms.
- (6) *Shallow outputs* – modules deliver output in the form of unanalysed representations.
- (7) *Specific neural architecture* – even though modules are mental rather than neural structures, they are often 'hardwired' in the brain; that is, they are implemented by the same neural structure in all normal brains.
- (8) *Characteristic breakdown patterns* – when modules are damaged or absent, this is manifested in a typical pattern of symptoms.
- (9) *Characteristic pace and sequencing in development* – modules are innately specified or genetically determined.

According to Fodor, a mental mechanism must have most or all of the nine properties listed above in order to qualify as a module.⁶ Indeed, it is the

⁵ I will not analyse these properties here, nor discuss the kinds of evidence that would count in assessing them. Such a discussion would involve a protracted detour and would risk obscuring the main line of argument. Besides, Fodor himself has already discussed each of these properties in detail as well as the methodological problems of investigating them – indeed, this discussion takes up most of his original book.

⁶ Dominic Murphy and Stephen Stich point out, the term 'module' is used in a much broader and less demanding way by most evolutionary psychologists. This different usage can lead to confusion, so Murphy and Stich suggest that in cases of ambiguity, the two concepts should be distinguished by using the term 'Fodorian module' to designate the stricter, original concept, and the term 'Darwinian module' to designate the broader notion employed by evolutionary psychologists (Murphy and Stich, 1998: 3).

regular co-occurrence of these nine properties that is supposed to make modularity a robust natural kind.

I have two claims to make about the modularity of emotion:

- (i) Firstly, an empirical claim: as a matter of fact, most of the things commonly identified as emotions in humans and other animals possess most of these nine properties. An overwhelming mass of psychological and neuroscientific research supports this claim, but to go into it in any detail would involve me in a protracted detour. Paul Griffiths sums up some of the evidence in his 1990 paper on psychoevolutionary theories of emotion (Griffiths, 1990). Empirically, then, emotions are properly described as modular.
- (ii) Secondly, a theoretical claim: the interruption theory predicts that emotions should have many of these properties. I have already argued that interruption mechanisms must have property (1): domain specificity. It should be pretty clear, without needing to spell out the argument, that interruption mechanisms must also have properties (2) and (4): they must be automatic and fast. In natural organisms, they should develop regularly in a wide range of environments – property (9). Property (7) seems to follow from this, and to lead naturally to property (8).

This point is not new. Paul Griffiths pointed out the links between evolutionary theories of emotion and Fodor's concept of modularity some years ago, first in a paper and then in a book, though he only thought it plausible that *basic* emotions were modular (Griffiths, 1990; 1997). More recently, however, Griffiths has disavowed his earlier views. He no longer thinks that it is useful to think of emotions as modular. In a recent article in *Metascience*, he proposes two reasons for this U-turn (Griffiths, 1999).

The first reason that Griffiths gives for rejecting modular accounts of emotion rests on the fact that modularity is supposed to be a whole cluster of properties, of which automaticity is only one. If these properties can dissociate relatively easily, as now seems to be the case, then it is wrong to call emotions 'modular' on the basis that they have one modular property, since this would prompt people to infer, falsely, that emotions must have all of the other modular properties too.

Griffiths's second reason for rejecting modular theories of emotion is that the concept of modularity involves a strong commitment to a crude form of nativism. The ninth criterion in Fodor's list states that

The Darwinian module is a broader concept than the Fodorian module because it does not insist that a mental mechanism possess all or even most of the nine properties listed above before calling it a 'module'. Murphy and Stich go so far as to claim that a mental mechanism need only possess properties (1), (2) and (5) in order to count as a Darwinian module, and sometimes even (2) and (5) are not even necessary.

modular mechanisms are innately specified by some kind of genetic and neural program. Griffiths had previously endorsed the application of this idea to basic emotions, but he now rejects it because he thinks that the concept of innateness involves 'the idea of a literal neural program, containing a representation of the species-typical behaviour which ensues when the program is *activated*', and this, he claims, is 'entirely the wrong way' to think about basic emotions (Griffiths, 1999: 50, emphasis in original).

Both of the reasons put forward by Griffiths for rejecting a modular account of basic emotions are rather foolish. The first reason misunderstands the nature of the claim about the modularity of emotion. If the claim is an empirical one, we are simply saying that emotions do, in fact, have most or all of the properties listed by Fodor. We are not claiming that one can *infer*, simply on the basis that emotions have one modular property, that they have them all. If the claim about modularity is a theoretical one, of course, we *are* saying that we infer that emotions have several of the properties in Fodor's list, but the inference is *not* from the possession of one modular property to the possession of the others. It is from a general theory about what emotions are – they are interruption mechanisms – to a set of empirical predictions about what properties we expect these mechanisms to have.

The second reason Griffiths gives for rejecting modular theories of emotion simply attacks a straw man. Just because we use terms like 'innately specified' and 'genetically determined' does not mean that we are committed to a crude form of nativism. These terms may be perfectly acceptable shorthand for a sophisticated view of development. Andrew Ariew, for example, has argued for an account of innateness in terms of C. H. Waddington's concept of canalisation (Ariew, 1996). On this account, to say that something is 'innate' is simply to say its development is buffered against a wide range of environmental and genetic variation.

Griffiths is wrong, then, to reject the modular accounts of emotion he previously endorsed. If anything, he should have *extended* his thesis and recognised that it is not just *basic* emotions that are modular.