# RETHINKING EMOTION

## NEW RESEARCH IN EMOTION AND RECENT DEBATES IN COGNITIVE SCIENCE

DYLAN EVANS

Department of Philosophy
London School of Economics and Political Science

Thesis submitted for the degree of
Doctor of Philosophy
(Ph.D.)

Submitted: 5 July 2000
Corrected version: 21 August 2000

# Abstract

Cognitive science is currently the scene of a number of exciting debates. The so-called 'classical' approach, which has dominated the field since the 1950s, is increasingly being challenged on various fronts. Evolutionary psychologists and researchers in artificial life accuse classical cognitive scientists of ignoring the fact that natural cognition is not designed to solve abstract problems and prove theorems but to solve particular adaptive problems. Those working with a 'situated' view of the mind are challenging the classical commitment to internalism. Finally, proponents of dynamical approaches claim that the discrete models favoured by the classical approach are too coarse-grained and impute too much internal structure to the mind.

In this thesis I argue that the 'non-classical' approaches are compatible with classical cognitive science, with the important proviso that compatibility comes in different kinds. In the final chapter I outline a vision of a comprehensive 'integrated non-classical cognitive science' that combines the three non-classical approaches into a single conceptual bundle.

I illustrate these claims about cognitive science in general with reference to a particular field of research: the emotions. Emotions were ignored by most classical cognitive scientists, though some models of emotion were developed within the classical framework. These models, however, provided no way of distinguishing emotion from cognition. I argue that the non-classical approaches remedy this problem, and together provide a new way of thinking about the emotions which I dub 'the interruption theory'. Since the interruption theory borrows insights from all three of the non-classical forms of cognitive science, it serves as a good example of the integrated non-classical approach that I recommend for cognitive science in general.

## *Contents*

# *Introduction:*

# *Varieties of cognitive science*

Cognitive science is currently the scene of a number of exciting debates. The so-called 'classical' approach, which has dominated cognitive science since the 1950s, is increasingly being challenged on various fronts. Evolutionary psychologists and researchers in artificial life accuse classical cognitive scientists of ignoring the fact that natural cognition is not designed to solve abstract problems and prove theorems but to solve particular adaptive problems. Those working with a 'situated' view of the mind are challenging the classical commitment to internalism. Finally, proponents of dynamical approaches claim that the discrete models favoured by the classical approach are too coarse-grained and impute too much internal structure to the mind.

Because of the challenges they pose to the classical approach, the evolutionary, situated and dynamical approaches may all be referred to as 'non-classical'. One of the main questions I address in this thesis is whether or not these non-classical approaches are compatible with classical cognitive science. I argue that they are, in fact, compatible with the classical approach, with the important proviso that compatibility comes in different kinds. In the final chapter I outline a vision of a comprehensive 'integrated non-classical cognitive science' that combines the three non-classical approaches into a single conceptual bundle.

It is hard to assess these sweeping claims about cognitive science in general without reference to a particular field of research. The emotions constitute one such field, and, moreover, one that is eminently suited to assessing the compatibility of the classical and non-classical approaches. Emotions were ignored by most classical cognitive scientists, and some of the main proponents of the classical approach even went so far as to claim that they were strictly beyond the purview of cognitive science altogether. Later, some models of emotion were developed within the classical framework, but these models provided no way of distinguishing emotion from cognition. I argue that the non-classical approaches remedy this problem, and together provide a new way of thinking about the emotions which I dub 'the interruption theory'. Since the interruption theory borrows insights from all three of the non-classical forms of cognitive science, it serves as a good example of the integrated non-classical approach that I recommend.

*What is cognitive science?*

Cognitive science is a massive field, grouping together many formerly distinct disciplines, such as artificial intelligence, linguistics and the neurosciences, with branches of philosophy, psychology, and anthropology. Indeed, it may be more appropriate to speak of 'the cognitive sciences' rather than of 'cognitive science' in the singular, since the diversity of theoretical approaches and methodologies that now refer to themselves as 'cognitive' may preclude any view of them as a single discipline. This is not my view. I think that the cognitive sciences have enough in common to warrant speaking of a single entity, 'cognitive science' in the singular, that has a 'classical' form and various 'non-classical' variants.

The two key features that are shared by all forms of cognitive science are:

(i) the computational theory of mind (CTM): the idea that the mind is a computer; and

(ii) a design-based approach: the methodological maxim that a good way to understand any natural mind is by designing artificial ones.

I think it is the second clause – the commitment to a design-based methodology – that most clearly distinguishes cognitive science from previous approaches to the study of the mind. The computational theory of mind is certainly important, but the idea of computation is such a loose notion that to define cognitive science on this basis alone would be to risk vacuity. It is not enough to say that the mind is a computer; one must then set out to think how such a computer might be built. When we succeed in building a thinking machine, we will know a lot more about thought. Similarly, when we are able to build machines that can feel happy or sad, we will know a lot more about emotion. This is what I mean by the cognitive approach to understanding emotions.

## *Varieties of cognitive science*

The two core features of cognitive science are shared by all the approaches that I discuss in this thesis. It is only this that allows such different kinds of approach to the study of the mind to be regarded as forms of *cognitive science* in particular rather than merely forms of *psychology* (which I take to be a much more general term). These two core features provide the unity underlying the different approaches.

Within cognitive science, thus defined, I distinguish different approaches based on the stance one takes on particular issues. The evolutionary approach is defined by its emphasis on functional questions; what is the mind, and its various components, *for*? In other words, why did the mind evolve? The situated approach is defined by its rejection of internalism. And the dynamical approach is defined by its preference for continuous models over discrete-state machines. The classical approach is defined, by default, by its obliviousness to evolutionary-functional questions, its commitment to internalism, and its preference for discrete-state machines.

During the first three decades of cognitive science, from about 1950 to about 1980, the characteristics that I take to define the *classical* approach went largely unchallenged. During that period, few cognitive scientists paid much attention to evolutionary questions, and most were committed to internalism and to modelling the mind with discrete-state machines. Nobody referred to this set of features as embodying a particular *form* of cognitive science. Many assumed that these features were just as essential to cognitive science as the commitment to CTM and to a design-based methodology. It was only when certain sections of the cognitive science community began to question these assumptions that it became clear that cognitive science was not, in fact, essentially committed to ignoring evolutionary questions, to internalism, and to discrete-state machines. Only then were these three features seen as defining a particular *form* of cognitive science, rather than defining cognitive science *per se.* The particular form of cognitive science that was marked by these three features came to be called the 'classical' approach in retrospect, when various dissident groups within the cognitive science community wished to question one of the features without thereby excluding themselves from cognitive science itself.

### *Pluralism*

The identification of particular *forms* of cognitive science had the merit of making clear that the bracketing of evolutionary questions, the commitment to internalism, and the preference for discrete-state machines were all logically independent from the basic idea of CTM and from the choice of a design-based methodology. However, it also had a downside; it fractured cognitive science into a set of warring schools, each of which had a tendency to exaggerate its disagreements with the others. A vision of the underlying unity of cognitive science was lost, and cognitive scientists used the new labels to pigeon-hole one another. To the proponents of the non-classical approaches, it became an easy rhetorical ploy to refer to everyone before 1980 as classical. A typology of approaches became a way of classifying scientists.

Scientists are people, however, and people are much more complex than schools of thought. The latter can be defined in conceptual terms, as I have done with the various forms of cognitive science, but people are not so consistent. It is rare to find a scientist who pursues one kind of approach with single-minded dedication throughout his whole life. Thankfully, most people are more flexible than that. There probably never was a pure classical cognitive scientist, in the sense of one who never said anything about evolution, nor ever doubted internalism, nor ever wondered about the possibility of modelling the mind in continuous terms.

Nevertheless, flexibility comes in degrees. Even though there may never have been a pure classical cognitive scientist in the sense just described, there have been, and still are, cognitive scientists who are more closely identified with one approach rather than another. One of the aims of this thesis is to persuade cognitive scientists to be more flexible. The integrated non-classical approach I recommend is all about such flexibility.

### *Alan Turing: the pioneer of integrated cognitive science*

In the minds of many cognitive scientists, the name of Alan Turing is associated exclusively with the classical form of cognitive science. Turing's work does not brim with references to evolution; his emphasis on internal memory seems to make him a strong internalist; and the 'universal machine' to which he gave his name was a discrete-state machine. All these features figure strongly in his great paper of 1950, 'Computing machinery and intelligence' (Turing, 1950), which can therefore be taken as inaugurating the discipline of cognitive science in general and its classical form in particular.

Turing's legacy, however, turns out to be much broader than this. When one looks more closely at the 1950 paper, for example, one finds there not just the seeds of the classical approach, but the germs of all the so-called non-classical approaches too. In the section on 'learning machines', for example, Turing proposes a method of designing machines based on an analogy with natural selection, thus anticipating the techniques of artificial life by almost forty years (Turing, 1950: 52). The distinction between memory and processing may be far closer to the situated approach to cognition, with its emphasis on exploiting external resources to ease the burden of computation, than is usually realised (Wells, 1998: 275). Turing’s remarks, in section about the importance of giving a computer a body in order for it to have the same experiences as a normal child anticipate current research in robotics (Turing, 1950: 53). And although Turing put his money on digital computers having sufficient resources to pass his test for machine intelligence, he did not argue that non-digital computers did *not* have such resources. Indeed, he clearly states that the human nervous system is 'certainly not a discrete state machine' (Turing, 1950: 47), and argues that digital machines could pass his test only because they are capable of mimicking the behaviour of non-digital systems sufficiently closely. It turns out that Turing anticipated many of the supposedly novel challenges to the classical approach that are generating so much debate in cognitive science today.

Turing's legacy is, then, far richer and more eclectic than is generally believed. It may be more accurate to regard him, not as the founding father of the classical approach, but as the first pioneer of the truly integrated cognitive science that I propose in my final chapter. Turing did not merely leave us with a fascinating thought-experiment about how to test machines for intelligence and a detailed set of proposals about one way (the classical way) to build such intelligent machines; he also left us with a number of provocative complementary suggestions, suggestions that have recently been developed independently by various cognitive scientists who object to the classical approach in one way or another.

The question of how Turing's rich legacy came to be reduced, in the minds of most commentators, to a mere fraction of what it really is, would make an interesting case study in the history of science. Whatever the reasons for this impoverished interpretation, one notes its subtle influence in even the most rigorous scholarship. When Douglas Hofstadter and Daniel Dennett, for example, published Turing's great paper in an anthology of philosophical works about the self, they chose to excise many of the passages I have just referred to as evidence of the richness of Turing's legacy (Hofstadter and Dennett, 1981).

J. A. Scott Kelso notes the same pattern of systematic misrepresentation in an interesting little aside in chapter one of his book, *Dynamic Patterns*. He tells a story about a famous scientist who was always arguing that the brain is not a Turing machine. When Kelso pointed out to the scientist that there was another Turing, the response was adamant: 'No, no, there's only one Turing. You know, the Turing of the Turing machine!' After Kelso wrote some equations on the board describing chemical patterns, the scientist paused and stared at him. 'Ah, I see what you mean,' he said. The equations that Kelso had written on the board were the very ones Turing had used to describe 'the chemical basis of morphogenesis' in another paper that has since become a classic in developmental biology (Turing, 1952; Kelso, 1995).

The anonymous scientist in this revealing anecdote had clearly heard of Turing's other work, for he recognised the reference when Kelso wrote up the diffusion equations that Turing had published in 1952. Yet, until he was reminded of this, the scientist insisted that there was 'only one Turing' – the father of the programmable digital computer. The 'other Turing' – the Turing who showed how patterns in nature can emerge without any programmer at all simply by means of a set of dynamic equations – had been eclipsed by the monolithic image of Turing the classical cognitive scientist.

Even Kelso is somewhat restrictive, however, when he captions his box, 'The two sides of Turing'. In addition to the two Turings that Kelso mentions, there are all the other facets that I have just mentioned, such as the Turing who anticipated recent ideas about the value of giving computers humanlike bodies, and the Turing who foresaw the field of artificial life. And it is not necessary to go to Turing's other papers to find these men. They are all there in the great paper of 1950.

Turing even seems to have anticipated recent ideas about the mind emerging from complexity theory. I do not examine complexity theory in this thesis, so I will conclude this introduction by a brief discussion of its relevance to cognitive science. Complexity theory takes ideas from dynamical systems theory and applies them to systems composed of a wide range of components or of a large number of similar components.[1] One of the key ideas in complexity theory is the idea of supercriticality. Complex systems usually exhibit one or more phase-transitions in which the addition of just a few extra components can produce an abrupt shift in the system's behaviour.

Turing's paper pre-dates the development of complexity theory by at least three decades, yet it contains a thought-provoking idea about the role of supercriticality in mental development. After discussing the phase transition in atomic piles that occurs when they reach critical mass, Turing asks if there is a corresponding phenomenon for minds. He answers in the affirmative:

> The majority of (human minds) seem to be 'subcritical', that is, to correspond in this analogy to piles of subcritical size. An idea presented to such a mind will on average give rise to less than one idea in reply. A smallish proportion are supercritical. An idea presented to such a mind may give rise to a whole 'theory' consisting of secondary, tertiary and more remote ideas. Animals' minds seem to be very definitely subcritical. Adhering to this analogy we ask, 'Can a machine be made to be supercritical?'
>
> (Turing, 1950: 51)

[1] There is no widely accepted definition of complexity theory. Some authors seem to treat it as a synonym of dynamical systems theory, or nonlinear dynamics, while others state explicitly that 'chaos is not complexity' (Bak, 1996). The central plank of the theory, in my view, is the idea that special mathematical tools are required for understanding systems in which the high number of components precludes any attempt to derive systemic properties directly from component properties. Care must also be taken to distinguish this theory from the branch of computational theory that is concerned with the complexity of certain mathematical functions, for this too goes by the name of complexity theory (Andy Wells, personal communication).

Another central idea in complexity theory is that complex adaptive systems tend to hover around the critical points, rather than living deep in the subcritical or the supercritical regions of their phase-spaces. In Stuart Kauffman's evocative terminology, complex adaptive systems tend to be poised 'at the edge of chaos', where behaviour is neither entirely ordered nor completely chaotic (Kauffman, 1995: 86-92). Kelso is somewhat more prosaic and prefers to speak of 'the intermittency mechanism', though his meaning is essentially the same (Kelso, 1995: 99). The following passage, again taken from Turing's 1950 paper, seems to anticipate just this idea:

> Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation, but a rather slight one, which does not give rise to random behaviour, or to pointless repetitive loops.
>
> (Turing, 1950: 55)

With hindsight and close-reading, then, Turing's prescience can be seen to extend even into the most recent thinking in cognitive science. Yet this does not match the common idea of Turing today as the founding father of the classical approach.

The impoverishment of Turing's legacy reflects a general narrowing of perspective that has marred cognitive science for much of its history. In this history, something has occurred akin to what Steven Jay Gould has called 'the hardening of the modern synthesis' in evolutionary biology (Gould, 1983). Varela, Thompson and Rosch argue that the cybernetics movement, which directly preceded the emergence of cognitive science proper, was characterised by a much more pluralistic approach to cognition – one that, for example, still regarded the digital nature of computation as an open question rather than an accepted dogma (Varela, Thompson et al., 1991: 37-39). If Varela *et. al.* are correct in their historical account, my arguments in chapter six about the need for a similar pluralism in cognitive science can be seen as a plea for cognitive science to return to its eclectic roots in what we might call the 'pre-classical' era of cognitive science in the 1940s, when the design-based approach was initiated by disciplines with other names such as 'control theory', 'information theory' and 'cybernetics'. The various species of cognitive science I discuss in this thesis – both classical and non-classical – can then be seen, not as antagonists, but as pieces in a complex jigsaw, all of which are necessary if we are to get the whole picture about mental phenomena. In the second part of chapter six, I go on to argue that this eclecticism is especially vital when it comes to getting the whole picture about emotion – although Turing, it must be said, was silent about this important part of our mental life.

# *Chapter One:*

# *Cognitive science and emotion*

> *The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent* without *emotions.*
>
> Marvin Minsky, *The Society of Mind*

In this chapter I discuss the two ideas which underlie all the different forms of cognitive science: the computational theory of mind, and the design-based approach. I then go on to show how these ideas also mark out the cognitive science of emotion as a distinctive way of understanding emotional phenomena.

## *1.1. Cognitive science*

Many commentators assume that the heart of cognitive science is the computational theory of mind (henceforth CTM) (e.g. Gardner, 1987). At first this idea seems quite appealing. Yet, as I argue in this section, CTM reduces to the old-fashioned representational theory of mind plus a commitment to materialism. Thus, if CTM were all there were to cognitive science, we would have to count almost all twentieth-century psychologists and neuroscientists as cognitive scientists. This would be stretching the term rather too far. Cognitive science is a distinctive approach to the study of the mind, and not all psychologists regard themselves as taking this approach. I conclude that, although cognitive science is indeed committed to CTM, this is not its most distinctive feature. The thing that really sets cognitive science apart from other ways of studying of the mind is the fact that it takes a *design-based* approach.

### *The computational theory of mind*

The term 'computer' originally referred to people who computed the answers to mathematical problems – that is, people who did sums. This, indeed, is how Alan Turing was using the term as late as 1950, by which time there were already a variety of electronic machines that could perform complex calculations automatically. These machines came to be known as *electronic computers* to distinguish them from their human counterparts. The etymology makes it clear that computers are defined in functional terms. So long as a thing can calculate the answers to certain mathematical questions, it is a computer, no matter what it is made of.

In claiming that minds are no more than 'things that compute', CTM makes a bold reductive move. CTM claims not merely that *some* minds are capable of performing mathematical calculations, but that *all* are, and that all the other things that minds do are reducible to such calculations.

This claim might have seemed rash in the early days of cognitive science, but as technology has progressed it has become more plausible. The range of things that electronic computers can do now is quite staggering, and yet all these capacities are achieved by means of reducing each step of each task to a set of mathematical operations. For a wide range of problems that were previously solvable only by agents with minds, we now have a transparent account of how they can be solved by machines that perform computations. *Prima facie*, this lends strong support to the idea that all *other* problems which are currently believed to require mental powers for their solution will eventually be solvable by computing machines.

Even if this is true, and all technical objections to CTM are removed, there may remain theoretical objections. Principal among these is the charge of vagueness that threatens the notion of computation. The basic idea of taking input and generating output in accordance with some mathematical function is so general that, if this were all that computation consisted of, practically anything could be construed as a computer. The position of all the bodies in the universe at time T2, for example, is a function of their position at time T1. If computation is just a question of systematically transforming input into output, we could regard the whole universe as a computer that takes the position of bodies at T1 as input and generates, as output, their position at T2. Yet it seems perverse to regard the universe as single gigantic mind. So, unless we can find some extra condition to constrain our notion of computation, this example would be enough to refute CTM.

Cognitive scientists generally argue that just such a constraint is provided by the notion of representation. According to this view, for x to be a computer, it is not enough that x systematically transforms input into output; the input-output relation must be representational. That is, the process that transforms input into output must be 'about', or designate, some process other than itself. Only when this is the case does it make sense to judge the output as being 'correct' or not. When the input-output transformation in x correlates well with another transformation elsewhere, the output of x can said to be correct. When the input-output transformation in x does not correlate well with another transformation elsewhere, the output x is incorrect. On this view, nothing is a computer in itself, but only with respect to some other system.

Once this constraint is imposed on the notion of computation, it no longer becomes possible to view the entire universe as one big computer. By definition, there can be no external system with which the changing positions of all material entities in the universe can be compared. One counter-example to CTM, at least, can be dispensed with. However, astronomers frequently run simulations of *parts* of the universe, such as the solar system. The machines running such simulations count as computers because there is a correlation between the way they transform input into output on the one hand, and the changing positions of the bodies in the part of the universe they represent on the other. However, since correlation is a symmetrical relation, it would be just as legitimate to regard the relevant *part of the universe* as a computer with respect to the simulation.[1] Yet it seems just as perverse to regard the solar system as a mind as to view the whole universe as one.

To defeat this counter-example to CTM, we need some way of introducing asymmetry into our notion of representation so that, whenever we have two systems, x and y, which transform input into output in accordance with similar functions, *one* system alone can be non-arbitrarily designated as the computer. We may be able to find a way of introducing such asymmetry by appealing to the idea of approximation. Astronomical models of the solar system are only ever *approximate*; that is why we call *them* simulations and regard the solar system as *the real thing.* The relation of approximation is non-symmetrical because the correlation between x and y can be increased by *adding* degrees of freedom to x and/or *subtracting* degrees of freedom from y, but not *vice-versa.*

Let us now pause to summarise the argument so far. According to CTM, having a mind just means being a computer, and anything that computes can be said to have a mind. Computers can be defined as systems that systematically transform input into output in a way that closely approximates the behaviour of some other external system. On this definition, the machines on which astronomers run simulations of the solar system count as computers. Hence, if CTM is right, such machines can be said to have minds.

This position has been dubbed 'strong AI' by the philosopher John Searle, to distinguish it from what he calls 'weak AI'. Weak AI is merely the idea that computers are powerful tools in psychology that enable us 'to formulate and test hypotheses in a more rigorous and precise fashion than before' (Searle, 1980: 183). Strong AI goes a lot further than this, claiming

[1] More generally, so long as computation is constrained only by the notion of representation, and representation is defined purely in terms of correlation, whenever there are two systems, x and y, which transform input into output in accordance with similar functions, *both* systems can be regarded as computers whose internal states represent those of the other.

that computers are not merely tools, but (when appropriately programmed) really have minds, and 'can literally be said to *understand* and have other cognitive states' (Searle, 1980: 183, emphasis in original). In strong AI, the programs do not merely help us to test psychological explanations; rather, the programs are themselves the explanations. That is, they are supposed to explain behaviour by providing precise models of the mental processes that generate it. Searle's term 'strong AI' is thus simply equivalent to the term 'cognitive science' as I use it in this thesis.

### *The representational theory of mind*

Constraining the notion of computation by appealing to the idea of representation strengthens CTM by excluding such obviously non-mental entities as the universe from the class of computers. It also ties CTM to an earlier tradition in the philosophy of mind. Representations are intentional – they are 'about' other things – and ever since Franz Brentano declared that intentionality is the distinguishing mark of the mental, there has been a thriving school of philosophical thought that identifies the mind with a set of representations (Brentano, 1874).[2] This is the so-called 'representational theory of mind' (RTM). Thus, by defining computers as representational systems, CTM seems to amount to no more than a re-statement of RTM.

CTM does, however, add something to RTM. By combining Brentano's thesis with the idea that the data in computing machines are *mental* representations, CTM was able to solve a problem that had beset earlier forms of RTM. Brentano and others in his wake had been accused of begging the question, since they offered no account of how mental

[2] More needs to be said, of course, about the way in which mental representations differ from, say, the linguistic representations on a page of print, or the pictorial representations in a painting, but CTM deals with this by appealing to the idea of *process*: minds are not just static sets of representations, but processes in which in representations are systematically transformed. To say that minds are processes does not imply, of course, that minds are not also metaphysically robust 'things'.

processes could be semantically coherent. That is, identifying thoughts with representations did not in itself explain how thoughts could follow each other in a way appropriate to their meanings. It seemed to some critics that Brentano's thesis merely pushed back the explanatory burden to some 'little man in the head', an inner homunculus who understood the *meanings* of the representations. It did not offer a clear account of how minds could be material entities.

By treating the mind as a computer, however, the first cognitive scientists argued that they could explain how thought processes are semantically coherent without positing such a homunculus. If all the rules for manipulating data are purely formal, based wholly on syntactic properties, and if these rules license all and only those inferences that are permissible on semantic grounds, then a commitment to mental representations can be compatible with a genuinely causal and materialistic account of the mind. There is no doubt about the purely physical make-up of computing machines, and such machines can be programmed to carry out the formal rules that respect the semantics of the symbols without recourse to a homunculus. I conclude that CTM is reducible to RTM plus a strongly argued case for materialism.

To sum up: according to CTM, minds are computers that process internal representations by means of purely formal rules. Mental processes, in other words, are determined by a program, which specifies how various symbolic representations are to be manipulated and transformed. The rules in the program, whether in a man-made computer or a human mind, are supposed to be precise, completely explicit, and exceptionless, so that an ability to perform elementary logical and mathematical operations is all that is needed to execute them. The individual components of the machine, therefore, can be quite 'dumb'; they need not 'understand' the content of the representations that the machine is manipulating, since they can treat the data and the rules as purely formal structures. The rules, then, apply to the representations purely on the basis of their formal syntactic structure, but because the syntax 'hangs together' with the semantics, the rules generate output that is properly interpretable as being about objects and facts in the external world.

*Criticisms of CTM*

Searle is famous for his criticisms of CTM. In his classic paper, 'Minds, brains, and programs', he argued that even the most appropriately programmed computer could never properly be said to have a mind because it could never *understand* anything (Searle, 1980). Searle claimed that computers were like a person who didn't understand Chinese, but who had a rulebook that enabled him to respond appropriately to whichever Chinese ideograms he was presented with. The Chinese room argument was intended to undermine the claim of classical cognitive science that computing machines are capable of having minds, but it only goes through if one accepts a number of assumptions. For example, the argument assumes that the computer is equivalent to the person in the room. This assumption has been challenged by various critics. The 'systems reply', which Searle discusses in his paper, argues that the computer is equivalent not to the person in the Chinese room, but to the whole system which comprises the room, the person, the rulebook, and everything else in the room. The person and the rulebook are analogous to the *components* of the computer. Just as 'understanding' is not ascribed to the individual components of the computer, but to the computer itself, so it is ascribed not to the person in the room but to the whole system. But the system is still representational because its answers may be judged as correct or incorrect by the external system constituted by the person *outside* the room.

I will not go into the various criticisms of the Chinese room argument and the various replies, which already constitute an ample literature by themselves. Suffice it to say here that there are still philosophers like Searle who do not find the claims of CTM convincing. Fodor may claim that CTM is 'the only game in town', but not all are persuaded. My own focus in this thesis, however, is not with the criticisms from outside cognitive science. Rather, I am concerned with the arguments of those who broadly *accept* CTM, but who wish to divorce it from other assumptions with which it is commonly linked. In the following chapters, I discuss various species of 'non-classical' cognitive science. While they may differ somewhat on how they define computation, and in their approach to hardware and software design, all of these species accept the basic idea that the appropriately programmed computer can be truly said to have a mind, and that the programs for these machines can themselves constitute *bona fide* psychological explanations.

### *Understanding by designing*

I have argued that CTM is reducible to RTM plus a well-argued case for a materialist view of mind. By appealing to the existence-proof of modern computing machines, CTM makes a good case for resolving important philosophical questions about how a commitment to mental representations can be compatible with a commitment to materialism. This, however, can hardly be used to pick out cognitive science as a distinctive research program in psychology. The vast majority of psychologists have, for over a century, adopted both RTM and a materialist view of mind. If CTM were all there were to cognitive science, the term would be rather vacuous.

If cognitive science is a distinctive research program, it must have some other feature peculiar to itself. I think that it does have such a distinguishing mark. In line with various other commentators, I take this to be its emphasis on taking a *design-based approach* (c.f. Haugeland, 1996) In other words, cognitive science is to be defined not simply by a theoretical commitment to the idea that the mind is a computer, but also by a methodological commitment to the idea that a good way to understand natural minds is by designing artificial ones.

This methodological maxim is intuitively very appealing. If you want to understand how a car works, one way might be to try and design a vehicle that exhibits similar properties. Likewise, argue cognitive scientists, if you want to know how the human mind works, one way to do this is to design an artificial mind that mimics the human mind at some acceptable level of similarity. As John Haugeland points out, this approach to understanding minds is rather different from traditional empirical psychology, which is often purely descriptive. Unlike traditional psychology, which works backwards from observable behaviour to hypothetical mental causes, cognitive science starts with a proposed mental design and then works forwards by constructing a machine along these lines and observing how its performance compares to that of a natural cognitive agent. If the performance is similar to some acceptable degree, then this is good grounds for thinking that the mind of the natural cognitive agent has a similar internal design to that of the machine. Haugeland coins the term 'mind design' to refer to this forward-facing methodology (Haugeland, 1996). The term nicely underlines the crucial role played by artificial intelligence and software engineering in the research program of cognitive science. The claim is not that learning to design a mind is the *only* way of doing psychology. Rather, the claim is that designing an artificial mind is a very good way of doing psychology that would, at the very least, complement other, more descriptive approaches.

Defining cognitive science by its commitment to a design-based approach places artificial intelligence at the core of cognitive science. This may annoy those cognitive scientists who are not actively engaged in building artificial minds, as it may seem to imply that their research is not as important as work in AI, or even that they are not 'true' cognitive scientists. This is not, however, the intended meaning of the second clause. The clause does not specify that all cognitive scientists must take an active part in *building* artificial minds. It simply states that cognitive scientists are those who adopt as a methodological maxim the idea that designing an artificial mind is a good way to understand natural ones. This condition is fulfilled, I claim, whenever researchers propose models of the mind that are computational enough to permit a computer program to be *readily* designed on the basis of the model. If a model of mental structure proposed by a psychologist, for example, is written in the form of a decision-tree or flowchart, this could easily be taken by a programmer and implemented on a computing machine.[3]

### *Machines and men*

The term 'machine' is often used by cognitive scientists as a convenient label for the hardware that is supposed to implement the artificial minds they design. However, as Turing pointed out, if the cognitive research program is not to become vacuous, we must be careful about how we understand this label. Machines are, by definition, artificial, yet it can be hard to draw a firm line between the artificial and the natural. Turing attempted to avoid getting bogged down in such tough metaphysical questions by simply stipulating that human beings born in the usual manner could not be regarded as true machines (Turing, 1950: 31). However, this stipulation is clearly not very stringent. Current advances in biotechnology make it conceivable that, in a few year's time, we may be able to clone a human being from a single adult cell and incubate the foetus in an artificial womb. The result would be a human being, but not one 'born in the usual manner'. It would, therefore, satisfy Turing's definition of the term 'machine'. Yet to claim that the resulting cognitive agent was a triumph for cognitive science would clearly violate the spirit of Turing's definition, if not the letter. Turing himself noted this possibility:

> ... it is probably possible to rear a complete individual from a single cell of the skin (say) of a man. To do so would be a feat of biological

[3] This is what happened with some of the work in appraisal theory that is mentioned in chapter one; some of these models were not written as computer programs, but they *were* written as decision trees, with more than just an eye to their potential implementation in a computer program. These models would, therefore, count as 'cognitive' on my definition. Likewise, many of the accounts of mental structure offered by evolutionary psychologists, and the accounts of neural structure provided by most neuroscientists, while not written as programs, are sufficiently computational in nature as to qualify as proper cognitive models. Most psychoanalytic models of the mind, on the other hand, would clearly be ruled out by the second clause.

> technique deserving of the very highest praise, but we would not be inclined to regard it as a case of 'constructing a thinking machine'.
>
> (Turing, 1950: 32)

*Transparent engineering*

To see why such things as a human being reared from a single somatic cell would not count as the realisation of the cognitive research program, we need not waste our time searching for more stringent definitions of the term 'machine'. Rather, we need to remember that the reason why cognitive science is interested in constructing machines with minds is in order to better understand the *natural* minds we observe around us. Software engineers may be content to build intelligent machines for practical purposes. So long as the machines can solve the problems they are built to solve, it will not worry the engineers if the machines seem to operate in ways that bear very little relation to the way natural minds work, or if the machines work in ways that are not fully understood. Cognitive scientists, however would not be content with such machines. Cognitive scientists require not only that their machines solve the kinds of problem that natural minds solve, but also that they do it in similar ways to those used by natural minds, and, furthermore, that the precise details of how the machines work are well understood. If we succeeded in constructing an artificial mind simply by mimicking the natural processes by which our brains develop, without understanding how the resultant construction operated, this would indeed be a stunning technical achievement, but it would not count as the culmination of the cognitive research program.

Thus it seems that cognitive science would only achieve its aim if it could build an artificial mind by means of a technique that is, to some extent, self-explanatory or *transparent*. By the phrase 'transparent engineering', I mean any method of construction whose principles are widely agreed by scientists to be well understood. Basic mechanical engineering is transparent in this sense, since we need only see the various pulleys, cogs and levers in a simple machine to understand how it works. 'Biological engineering', which is how we might describe the technique of growing a neural network in a petri dish, is not self-explanatory, since we still want further explanations of how neurons actually work in terms of simple mechanics. The reason why constructing silicon-based minds may be more informative, at the current moment, than constructing neuron-based minds is that the former are well understood in terms of their physical properties, whereas the latter are much more complex. One of the most puzzling things about minds is how properties such as intelligence and intentionality can arise from arrangements of mere matter. If we are seeking to understand how this occurs, it is surely better to work with materials whose physical properties are well understood. Otherwise, we risk begging the question.

### *Functionalism and multiple realisability*

The requirement that cognitive scientists build their machines out of components whose physical properties are well understood highlights an important feature of cognitive science – namely, the extent to which it is predicated on the assumption that 'mind' is a substrate neutral concept. If minds were tied to the neural tissue in which they are instantiated in humans and other vertebrates so intimately that they simply could not be instantiated in any other material, the whole edifice of classical cognitive science would come tumbling down. The idea that minds can be instantiated in many different media is known as the 'multiple realisability thesis', and this thesis is one of the cornerstones of the whole cognitive research program.

In its strongest form, the multiple realisability thesis implies that there are only the very weakest material constraints on the instantiation of any kind of functional organisation. It was this intuition that allowed Hilary Putnam and others to challenge the (type-type) identity theory of mind in the 1960s. Since minds, they claimed, are defined entirely in functional terms, and since we can imagine the same functions being performed by very different kinds of physical structure, it seems chauvinistic to deny that alien life forms and robots could have minds just because they do not have human-like brains (Putnam, 1960).

It is important to recognise, however, that the multiple realisability thesis is, at the moment, not proven. At present we have some evidence from artificial intelligence that minds like ours can be implemented by very different material structures, but this is not conclusive. It may well be the case, as some now argue, that human-like minds are much more dependent on the particular physical and biochemical properties of vertebrate neurons than previously thought. The multiple realisability thesis should be treated as empirical matter requiring further investigation, and not assumed on the basis of stories of robots and aliens that are, at present, mere science fiction. Indeed, testing the multiple realisability thesis can be seen as one of the subsidiary goals of artificial intelligence.

The multiple realisability thesis has led many classical cognitive scientists to take a very dismissive attitude towards neuroscience. The study of the brain becomes of very little interest once the mind is regarded as software which can, in principle at least, be run on almost any kind of hardware. The cognitive psychologist can then elaborate hypotheses about the programs instantiated in the human brain without worrying at all about *how* these programs are instantiated. True, it might also be interesting to know about the details of instantiation, but this information could not provide any constraints on the development of hypotheses about the purely functional mechanisms studied by the cognitive psychologist, since these are substrate neutral. This has led some critics to accuse cognitive science, and the functionalist doctrine on which it is based, of a hidden Cartesianism (Edelman, 1992).

Daniel Dennett argues that these criticisms are misplaced because they fail to distinguish between two claims. The first is the broad idea of functionalism; the second is a specific set of minimalist empirical wagers (neuroanatomy doesn't matter, neurochemistry doesn't matter, etc.). It was the second claim, not the first, that provided an excuse for many early cognitive scientists to remain in blissful ignorance of neuroscience. In the past few decades, as it has become increasingly clear that the neurobiological details *do* matter, cognitive scientists have had to give up on their minimalist wagers and get to grips with nueronanatomy, neurochemistry and the rest of neuroscience. This has left the mistaken impression in some places that the underlying idea of functionalism is flawed. In fact, however, the correct inference to draw from these recent discoveries is precisely the opposite; the reasons why the new claims matter is precisely because we *accept* the broad idea of functionalism. As Dennett remarks, 'neurochemistry matters because – and *only* because – we have discovered that the many different neuromodulators and other chemical messengers that diffuse through the brain have *functional roles* that make important differences' (Dennett, 1999, author's manuscript, emphasis in original). What recent discoveries about the importance of neurobiology show is simply that functionalism has to be expanded downwards to include the details of the brain. Human minds may well be computers, but not the relatively simple computers that the first cognitive scientists hoped they would be. Their computational resources reach down into the sub-cellular level, and artificial minds that have humanlike intelligence will have to employ virtual neuromodulators and other such software that mimics these molecular resources.

Exactly how far downwards the computational resources of the human mind reach is a moot point. Roger Penrose has suggested that they reach as far down as the subatomic level. He argues that consciousness, in particular, depends crucially on the quantum mechanical properties of microtubules that are found in vertebrate neurons (Penrose, 1989). This is an extremely speculative claim, with no real evidence to back it up, and is regarded by many cognitive scientists with a considerable degree of scepticism. Nevertheless, if Penrose were correct in supposing that some subset of cognitive processes could only be realised by structures with particular quantum mechanical properties, the multiple realisability hypothesis would be severely weakened. Unless all physical materials had the relevant quantum-mechanical properties, there would be strong constraints, of a purely physical nature, on the kind of materials from which conscious minds could be constructed. Silicon, for example, might have inherent limitations which rule it out as a substrate for complex minds. This is, of course, an empirical matter that will only be resolved with further research in artificial intelligence.

### *Functional decomposition*

The way that functional hypotheses about the structure of the mind are extended downwards into the neurobiological and physical details is usually by means of an explanatory strategy known as 'functional decomposition'. This strategy works well for understanding how complex man-made machines like cars and computers work. It consists of picking out the various components from which the machine is made, described in terms of the functional role they play. For example, in a car we can identify various systems such as the ignition system and the combustion system. We can then proceed in the same way with each of these systems, identifying the various functional subsystems which compose them.

This strategy of breaking complex systems down into their components and subcomponents is often applied to natural biological systems as well as to man-made artefacts. In describing the physiology of an animal, for example, it is common to proceed by identifying various systems such as the endocrine system and the nervous system. These systems can then be further analysed. For example, we can take the nervous system and break it down into the sympathetic and the parasympathetic nervous systems. This strategy has worked well in biology despite the occasional objection to the apparent literalness with which it takes the analogy between organisms and artefacts. Largely because of its success, it has been taken as providing a model for psychological explanation by many cognitive scientists.

When applied to the mind, the strategy of functional decomposition is sometimes known as 'homuncular functionalism'. The idea here is that the mind can be broken down into various functional units, each of which can be imagined as a 'little man in the head' or homunculus. Each of the functional units can then be further analysed into subunits, which can be compared to even smaller mini-homunculi in the heads of the first homunculi. Unlike the traditional form of homunculism, according to which the man in the head was just as clever as the man in whose head he sat, homuncular functionalism is supposed to block infinite regress, and thus avoid vacuity, by requiring that the homunculi posited at each stage of the analysis are dumber than the homunculi posited at the previous stage (Fodor, 1968). Eventually, it is supposed, we will reach a stage at which the homunculi are so dumb as to be virtually mindless. That is, our psychological explanation ends when it is able to analyse a mini-mini-mini homunculus into components that can be understood in transparently mechanical (or neural) terms, without the need for any mentalistic or intentional vocabulary.

Of course, the talk of 'little men in the head' is merely a way of making the explanatory strategy more vivid. The strategy can be described in equivalent but less anthropomorphic terms by reference to the idea of a computer flowchart. Instead of being compared to little men, the functional units of the mind may be compared to the boxes in a computer flow chart. In such a flow chart, every box name is the name of a problem. 'If the computer is to simulate behaviour, every box name will be the name of a psychological problem' (Fodor, 1968: 48). Each box in the flowchart can then be analysed as a flowchart in its own right, and so on, until the boxes in the last flowcharts are clearly realisable by transparent engineering.

This completes our brief survey of cognitive science. In the following section, I show how the basic principles of cognitive science can be applied to the study of emotion.

### *1.2. The cognitive science of emotion*

Insofar as the study of emotion is concerned, the choice of the word 'cognitive' to denote a research tradition in psychology was a recipe for misunderstanding. Many psychologists use the term 'cognitive' to refer to 'unemotional' thought processes, such as the deductive reasoning one might engage in when in a calm frame of mind. Yet, as I argued in the last section, when used to refer to a distinctive approach to the study of the mind, the term 'cognitive' means something quite different. There are, in other words, at least two quite different meanings of the term:

(1) When used to describe a way of studying the mind, as in the phrase 'cognitive science', the term denotes an approach that both (i) is committed to CTM and (ii) adopts a design-based methodology.

(2) When used to describe a mental faculty that contrasts with the emotions, the term denotes a set of mental processes whose paradigmatic forms are deductive reasoning, decision-making and problem-solving.

The fact that the word 'cognitive' can have both of these meanings has at times muddled the debate about emotion in cognitive science. For example, it can lead to the false impression that cognitive science must be concerned exclusively with understanding unemotional thought processes.

The first generation of cognitive scientists were largely of this view. In his classic textbook, *Cognitive Psychology*, Ulric Neisser stated unequivocally that dynamic and motivational factors such as emotions were not part of the field (Neisser, 1967). Jerry Fodor echoed this view in *The Language of Thought* (Fodor, 1975), and Howard Gardner has listed the de-emphasis of affective or emotional factors among five defining features of cognitive science (Gardner, 1987).

When one looks at the kind of programs written in the first decades of cognitive science, the exclusion of emotional processes is strikingly obvious. From the chess-playing programs to the theorem-provers, none seems to exhibit any feature that even remotely resembles an emotion. On the contrary, they all model our paradigmatic notions of unemotional thought processes. Computers running such programs behave like high-functioning autistics. Like these so-called *idiots savants,* machines running early classical programs are unusually gifted in certain areas, such as 'rapid computation of large numbers, memorising phone listings, and precise memory of huge sets of facts and trivia, but ... lack the forms of common sense and emotional intelligence that most people acquire effortlessly' (Picard, 1997: 90).

However, while it is certainly true that the first models of the mind that were cognitive in sense (1) did, in fact, happen to deal exclusively with mental processes that were cognitive in sense (2), this need not have been the case. The two sense of the word cognitive are logically independent, so there is no contradiction involved in speaking of the cognitive science of emotion. To think that there is would be to confuse the two senses of the word cognitive.

There is nothing that rules out taking a design-based approach to emotion as well as to cognition. The cognitive approach to cognition is based on the idea that, by attempting to design machines that can think, we will come to know a lot more about thought. The cognitive approach to emotion is based on the idea that, by attempting to design machines that emote, we will come to know a lot more about emotion.

### *Reason and the passions*

All this is to presuppose that cognition and emotion (or, in an older vocabulary, reason and the passions) are distinct types of mental process. This idea is an old one, going back at least as far as Plato, but precise nature of the distinction is hard to pin down. In the eighteenth century, David Hume attempted to specify exactly how reason differed from passion by appealing to the idea of representation. In *A Treatise on Human Nature*, he argued that reason was representational while the passions were not. On this view, thoughts can be judged as true or false, while emotions cannot:

> A passion is an original existence, or, if you will, modification of existence, and contains not any representative quality, which renders it a copy of any other existence or modification. When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high. 'Tis impossible, therefore, that this passion can be oppos'd by, or be contradictory to truth, or reason; since this contradiction consists in the disagreement of ideas, considered as copies, with those objects, which they represent ... nothing can be contrary to truth or reason, except what has a reference to it, and ... the judgements of our understanding only have this reference...
>
> (Hume, 1734: 415)

Hume's thesis about the non-representational nature of emotion has been very influential, but it poses an obvious dilemma for cognitive science. As we saw in the previous section, cognitive science subscribes to CTM, which is a version of the representational theory of mind. So cogntive science must either reject Hume's thesis or exclude emotions from the class of mental processes. In the next chapter, I outline the way in which some of the first cognitive scientists responded to this dilemma by constructing a representational account of emotion.

### *The substrate neutrality of emotion*

Before concluding this chapter, however, it is worth noting that the multiple realisability thesis applies just as much to emotion as to other kinds of mental process. In other words, one we are committed to a view of emotion as a computational process and to CTM, then we must conclude that emotions are to be defined in functional terms rather than in purely material ones. Emotions, that is, are substrate neutral.

This means that we cannot refuse to attribute true emotions to a machine simply on the grounds that it is not made out of flesh and blood. This, however, may be harder for many people to accept than the corresponding idea that machines could truly be said to think even if they are made of different materials. Cultural representations of intelligent machines abound, but such machines are generally devoid of emotion. In many people's minds, emotions are precisely what distinguish us from machines (Turkle, 1984). The cognitive science of emotion rejects this view as too anthropomorphic. So long as machines have components that perform the same function as the neural mechanisms of emotion in humans, the machines could be said to have true emotions.

# *Chapter Two:*

# *Classical cognitive science and emotion*

> *Think how much stronger the self will be when it deliberately uses reason and judgement to form a decision. For the mind freed from passions is like a fortress, and there is nothing more secure in which to retreat and find unceasing sanctuary.*
>
> Marcus Aurelius, *The Meditations*

It may seem presumptuous that a discipline barely forty years old already prides itself on having a 'classical' form and various 'non-classical' variants. Yet this is exactly how cognitive scientists describe the theoretical diversity that currently characterises their field of study. In this chapter I outline the main features of classical cognitive science, and then discuss the classical approach to emotion.

## *2.1. Classical cognitive science*

By 'classical' cognitive science, I intend to refer to the general style in which the cognitive research program was pursued in its first decades, from about 1950 to 1980. This style was marked by a number of assumptions that have since been challenged by various sections of the cognitive science community. These assumptions included:

*(i)* *Domain generality*
*(ii)* *Internalism*
(iii) *Discreteness*

In the rest of this section, I will briefly describe what is meant by each of these terms. First, however, I want to make a few general points.

During the years 1950-1980, nobody spoke of 'classical' cognitive science. The assumptions of domain generality, internalism and digitality were so widely accepted by cognitive scientists that they seemed essential components of cognitive science itself. Only later, when various sections of the cognitive science community began to challenge these assumptions, did it become clear that that cognitive science was not, in fact, committed to them. At that point, it became useful to distinguish different *species* of cognitive science. Because the assumptions of domain generality, internalism and digitality had prevailed in the early days of cognitive science, it became common to refer to this set of views as marking the 'classical' form of the discipline. Those who challenged one or more of these assumptions could then refer their own approaches as 'non-classical'.

The classical assumptions of domain-generality, internalism and digitality were challenged, respectively, by those adopting evolutionary, situated, and analogue approaches to the study of the mind. These approaches therefore, are all 'non-classical' in one way or another. However, just because an approach is non-classical in one way does not mean that it has to be non-classical in every other way. Just because the evolutionary approach rejects domain-generality, for example, does not mean that it has to reject internalism and digitality. However, in the final chapter I argue that while the various non-classical approaches are *logically* independent of one another, there are good *theoretical* reasons of a non-logical kind that make the non-classical approaches natural allies. While they do not logically entail one another, the non-classical approaches can be combined to make up a single coherent species of cognitive science that we might refer to as 'integrated non-classical cognitive science'.

This said, it is now time to look at what the assumptions of classical cognitive science actually say.

### *(i) Domain generality*

As I noted in the previous chapter, the principle explanatory strategy of cognitive science is functional decomposition. On this view, understanding the mind is like understanding the body in the sense that both involve 'carving nature at its joints'. Cognitive science is supposed to proceed by producing an 'anatomy' (or a 'map') of the human mind, whose components are discovered by taking a componential approach to designing artificial minds. If the artificial minds thus designed behave just like human ones, this is good grounds for assuming that the human mind is composed of similar functional units.

In adopting this view, cognitive science assumes that the mind can be partitioned into distinct parts. This view, which is sometimes known as 'faculty psychology' is an ancient view, going back at least as far as Plato.[1] There are various ways in which one could go about anatomising the mind, but cognitive scientists in the period 1950-1980 assumed that the best way of doing so was to divide the mind, first, into three parts: a perception system, a general reasoning system, and a motor-control system. This was not a new idea either; it figured in many previous models of the mind, including one proposed by Freud.

In the nineteenth century, however, Franz Joseph Gall had proposed a rather different way of doing mental anatomy. According to Gall, the mind is composed of a large number of subsystems, each of which operates solely in a particular *domain*. Gall did not spell out in any detail the criteria for individuating domains, but one gets a rough intuitive idea of what he meant when one reads the list of subsystems he identified; there were, for example, subsystems for musical ability, moral reasoning, and the appreciation of beauty (see Fodor, 1983).

[1] Faculty psychology was challenged in the eighteenth century by Hume and other associationists, who argued that the mind is a homogeneous entity governed by a few general principles. If the associationists are right, then the viability of functional decomposition as an explanatory strategy in understanding the mind would be seriously in doubt

Gall's taxonomy of mental structures is clearly orthogonal to the taxonomy assumed by most early cognitive scientists. In that taxonomy, a single perceptual system took in all sensory stimuli, processed them, and passed them to a single reasoning system, which then decided, on the basis of all this information, what instructions to pass on to the motor control system. There is no specific sensory system dedicated purely to the analysis of music or beauty. The sensory system, like the reasoning system and the motor-control system, is *domain-general*.

In Gall's view, things are quite different. Each subsystem works relatively autonomously with input of a certain class. The music subsystem only processes musical stimuli, for example. Each subsystem has its own relatively independent means for perceiving, reasoning and initiating motion. Another way of putting this is that perception, reason and motor-control are *domain-specific*.

The choice between the two taxonomies is not a black and white one. In 1983, for example, Jerry Fodor proposed a hybrid taxonomy that combined elements of both. In *The Modularity of Mind*, he argued that the sensory systems and motor-control systems were domain-specific, but that the central reasoning system was domain-general (Fodor, 1983). Fodor's criteria for individuating domains, however, were rather different from those envisaged by Gall. In fact, they were just the same as those normally used to individuate the five senses (with the modification that language comprehension was regarded as a distinct perceptual system in its own right, a kind of 'sixth sense'). According to Fodor, the distinct sensory systems worked relatively autonomously, but then passed all their information to a single, domain-general central system where all the data were integrated. Fodor's mind is still, then, fundamentally domain-general in its centre and bulk. Despite the small gesture towards domain-specificity, Fodor retains the basic assumption of domain-generality that characterises the classical approach.

Fodor's suggestion opened the way for other cognitive scientists to argue for a view much more like Gall's. In particular, a number of evolutionary psychologists began to argue, in the late 1980's, that the human mind was composed entirely of domain-specific systems, each of which had evolved to solve a particular adaptive problem faced by our ancestors. This has since become known as the 'massive modularity hypothesis'.[2] I discuss this view in chapter three.

### *(ii) Internalism*

CTM states that minds are computers, but (as we saw in chapter one) nothing is a computer except with regard to some other, external system. Whenever we are confronted with a claim that something has a mind, therefore, we can always ask where, exactly, the boundary lies between the internal and the external systems. Where, in other words, are we to locate the input-output boundary?

In the case of humans and other brainy creatures, most cognitive scientists in the period 1950-1980 tended to locate this boundary at the junction between the central nervous system and the rest of the body. The mind, in other words, was thought to *supervene* entirely on the brain. The only *bona fide* psychological states, therefore, are those that can be individuated without reference to the rest of the world outside the brain. This view is sometimes known as internalism or individualism, though, like most 'isms', these words cover a multitude of sins and mean different things to different people. *The MIT Encyclopedia of the Cognitive Sciences*, for example, takes a slightly different view of internalism, defining it as the view that 'psychology in particular and the cognitive sciences more generally are to be concerned with natural kinds whose instances end at the boundary of the individual' (Wilson, 1999: 397). Equating the input-output boundary with that of the *individual* is clearly somewhat different to equating it with the *brain-body* boundary.

[2] The term is due to Sperber (1994).

Instead of trying to resolve this ambiguity by futile discussions about the where exactly internalists *should* locate the input-output boundary, we can simply identify the common assumption underlying all the different formulations of internalism. Although they may disagree about *where* the input-output boundary is located, all internalists assume that this boundary is to be *identified* with some physical feature of the organism. It is *this* assumption, therefore, that should be taken as the essence of internalism.

In the 1980s, a loose federation of cognitive scientists began to reject internalism in favour of a more 'situated' approach to the mind. They argued that the input-output boundary was a moveable feast. On their view, the question of where the input-output boundary is to be located will depend very much on the particular context of enquiry. The boundary of the *mind* is not to be simply *identified* with *any* physical boundary, whether that between the brain and the rest of the body or that between the body and the rest of the world. Some mental processes may well supervene entirely on the brain, or even on just one part of the brain. Others supervene on the brain plus part of the body. Others supervene on the brain, body, and parts of the world. In Andy Clark's arresting metaphor, the mind often *leaks* out of the brain into the body and the rest of the world (Clark, 1997). I discuss this view in chapter four.

### *(iii) Discreteness*

The machines designed by cognitive scientists in the period 1950-1980 were almost all *discrete-state* machines. In such machines, the transitions from one state to another are like sudden jumps or clicks. The various states are sufficiently distinct and definite for the possibility of confusion between them to be ignored. There are no intermediate positions between one state and another.

Many of the first generation of cognitive scientists assumed that this was the *only* way to design intelligent machines. In 1976 Allen Newell and

Herbert Simon consolidated this impression by arguing that all cognitive agents would turn out to be digital machines. This is the thrust of their 'physical symbol system' hypothesis (Newell and Simon, 1976: 85). Strictly speaking, this claim has nothing to do with discreteness. A digital machine is not *necessarily* a discrete-state machine. A digital machine is simply one that can reidentify things it has made positively and reliably (Haugeland, 1996: 9).

Now, it *might* turn out to be the case that only discrete-state machines can succeed in being digital, but, if so, this would be an empirical discovery about discrete-state machines – we cannot make any such inference on purely conceptual grounds. The concepts of being discrete and being digital are logically independent. Nevertheless, for some reason the two concepts have often been confused, with the result that Newell and Simon's physical symbol system hypothesis seemed to confirm the view that cognitive science was exclusively concerned with discrete-state machines.

In the 1980s, this view was challenged by a growing band of cognitive scientists who were interested in using *continuous* machines to model the mind. They used components with continuously variable rates of activation, and linked them together in networks resembling groups of interconnected neurons. The connectionist movement was not the only section of the cognitive science community to call for analogue models of the mind. In the 1990s, they were joined by others of a more theoretical bent who proposed that cognitive science could benefit from using the tools of dynamical systems theory.

Dynamical systems are not all continuous. There are discrete dynamical systems as well. But most proponents of dynamical approaches to cognition have tended to concentrate on continuous systems. It is therefore plausible to regard the connectionist movement and the dynamical approach to cognitive science as fighting on a similar front. Both call into question the assumption that cognitive science is exclusively concerned with discrete-state machines. For this reason, I discuss them both together in chapter five, which is concerned with continuous approaches to cognitive science.

### *Other features of classical cognitive science*

A number of other assumptions were shared by many of the first cognitive scientists apart from the three listed above. Many, for example, assumed that mental representations are stored in the brain in a rich language-like code (dubbed 'Mentalese' by Fodor) which is independent of any natural language like English or Japanese (Fodor, 1975). This is known as the 'language of thought' hypothesis', and is a particularly strong form of representationalism. It is possible to adopt weaker forms of representationalism, according to which thoughts are representations, but do not take a language-like, propositional form. Cognitive scientists working outside the classical tradition are generally more likely to espouse such a weaker form of representationalism, and a few even claim to reject the idea that thoughts are representations altogether (although it is not clear whether, in fact, they are really just objecting to the strong Fodorian version). One can then, with hindsight, see a commitment to the language of thought hypothesis as another of the distinguishing features of classical cognitive science.

I take it, however, that the language of thought hypothesis is less central to the classical approach than the three assumptions described above. Whole movements in cognitive science have arisen based on the desire to challenge these assumptions. I think, then, that there are good grounds for considering these assumptions to be the main diagnostic features of the classical approach. In the next section, I show how they inform the classical approach to emotion.

## 2.2. The classical approach to emotion

Despite the general reluctance to address the emotions among most of the pioneers of cognitive science, there were, even in the early days, a few lonely voices calling for a more inclusive research program that would embrace emotional factors as well as classical cognitive processes. In 1963, for example, Robert Abelson proposed that cognitive psychologists should move away from their focus on 'cold' logical processes and address 'hot cognitions' (Abelson, 1963). In 1967 Herbert Simon himself argued cognitive models should include emotions (Simon, 1967).

Such calls for a more inclusive research program forced cognitive scientists to face up to the dilemma posed by Hume's distinction between reason and the passions. Since nobody was prepared to argue that emotions were not *bona fide* mental processes, cognitive scientists were forced to reject Hume's thesis about the non-representational nature of emotion. Hence, those cognitive scientists interested in emotion attempted to reduce emotions to particular kinds of thought. The resulting research project came to be known as appraisal theory.

### *Appraisal theory*

Appraisal theory assumes that emotions are evaluations of current situations. The first proponent of this approach was Magda Arnold, whose pioneering book, *Emotion and Personality*, practically inaugurated the cognitive science of emotion (Arnold, 1960). The research program that grew out of Arnold's work attempted to discover the features of situations and events which cognitive agents use to arrive at an emotional evaluation. For example, some have suggested that a key aspect of the antecedent situation is whether it was self-caused or other-caused (Smith and Ellsworth, 1985). On the basis of this and other criteria, represented in terms of a simple decision-tree, a computer could analyse any situation (if presented appropriately) and decide which emotion it should respond with. That is, it could take a linguistic description of a situation as input, and generate the name of a particular emotion as output. Such a computer would clearly have explicit internal representations of emotions.

During the past decade, an increasing number of computer models of emotion have been designed along these lines. Most of them have been based on two appraisal-type theories of emotion: the Ortony Clore Collins model (henceforth OCC) and Ira Roseman's model (Ortony, Clore et al., 1988; Roseman, Antoniou et al., 1996). Both of these models were constructed with computers in mind, so both are relatively easy to implement in software. The process of designing such simulations can help to test the theories as well as stimulating new questions (Picard, 1997: 195).

Both the OCC model and Roseman's model categorise emotions on the basis of the cognitive appraisal that people make about eliciting conditions. In the OCC model, emotions arise from valenced reactions to situations consisting of the consequences of events (is it good or bad for me and for others?), the actions of agents (do I approve of my actions and those of others?), and the aspects of objects (do I like them or not?). A decision tree involving these questions leads to twenty-two different emotions, from joy and distress to gratitude and anger. Roseman's model uses slightly different parameters to generate a total of seventeen emotions.

Both of these systems are framed in terms of rules and so are relatively easy to implement in classical digital computers (symbol systems). The OCC model has never been implemented in full in any AI system, but simplified versions of it have been used to synthesise emotions in various applications. Tomoko Koda used the OCC model to simulate a restricted set of ten emotional facial expressions on poker-playing software agents (Koda, 1996, cited in Picard, 1997). Clark Elliot augmented the OCC model to twenty-six emotion types and used these as the basis of his 'Affective Reasoner' system (Elliot, 1994).

## *The classical approach to emotion*

Within the field of cognitive science, appraisal theory may rightly be called the 'classical' approach to emotion, both because it has dominated the cognitive psychology of emotion, and because it implicitly adopts all the main tenets of classical cognitive science. In line with classical cognitive science, appraisal theory takes an implicitly domain-general, internalist and discrete view of emotion.

### *(i) Domain-generality and emotion*

In appraisal theory, there is no provision for distinct emotional subsystems. A single system analyses all the relevant features of the situation, and then computes a single emotion as output.

### *(ii) Internalism and emotion*

Appraisal theorists assumed that emotional processes supervened entirely on the brain. The brain required sensory and proprioceptive inputs, of course, in order to generate an emotion, and then instructed the body to move in certain ways in accordance with the kind of emotion generated, but in their 'essence', emotional processes were neural, not physiological.

### *(iii) Discreteness and emotion*

The decision-tree model of emotion employed by appraisal theory implies that emotions are discrete in a number of ways. Firstly, an emotion is either present or it is not; there are no intermediate levels of activation. Secondly, no attention is paid to the temporal features of emotion, such as how long it lasts, and how quickly it is triggered.

### *Criticisms of the classical approach to emotion*

The classical cognitive approach to emotion rejected Hume's explication of the traditional distinction between reason and the passions, but failed to specify any other way of explicating this distinction. In so doing, the classical approach to emotion threatened to undermine the distinction altogether and thereby eliminate emotion from the taxonomy of mental processes. It was not long before critical voices were raised. One of the most prominent critics was the psychologist, Robert Zajonc, whose influential article 'Feeling and thinking: preferences need no inferences', argued that it was important to retain the ancient distinction between cognition and emotion (Zajonc, 1980).

Zajonc's arguments, however, were weakened by a serious equivocation; his use of the word *cognitive* was decidedly ambiguous. At some points in his article, the term is used to designate a set of mental processes that differ in important ways from other mental processes of an 'emotional' nature. However, the experimental evidence that Zajonc cites in support of this distinction in fact supports a rather *different* claim; namely, that the appraisal process preceding the experience of emotion is largely inaccessible to *conscious* introspection (Zajonc, 1980). Thus the word 'cognitive' is best construed, at other points in the article, as synonymous with the term 'conscious'. This is misleading, to say the least. Unless we simply wish to identify conscious/unconscious distinction with the cognition/emotion distinction (which seems a very unattractive option), we must acknowledge that Zajonc's experiment does not provide grounds for regarding cognition and emotion as distinct kinds of mental process.

Zajonc's article was based, in part, on a clever experiment that he conducted himself, which was an extension of earlier work he had done on the 'mere exposure effect'. This term refers to the fact that, when subjects are exposed to novel visual patterns, and then asked to choose whether they prefer these or similar patterns to which they have *not* been exposed, they prefer the pre-exposed ones. Mere exposure, in other words, is enough to create preferences. In the experiment described in the 1980 paper, Zajonc presented the visual patterns so quickly that subjects were unable to state accurately whether or not they had seen them before. All the same, the mere exposure effect was still there. Subjects gave all sorts of reasons for preferring the pre-exposed patterns, but nobody gave the correct reason (the pre-exposure).

This experiment dealt a severe blow to the method then used by appraisal theorists to investigate the emotions. This method consisted of asking people to introspect and figure out what had gone through their minds just before they had had some emotional experience. The appraisal theorists hoped thereby to find out the rules that relate antecedent situations to consequent emotions. Zajonc's experiment showed that this method was flawed because people were often wrong about the mental processes that caused them to have emotions.

Zajonc *could* have concluded that the appraisal processes preceding the experience of emotion are often inaccessible to conscious introspection. However, he went much further than this, claiming that emotional preferences could be formed without the aid of any cognitive processes at all. This is clearly a *non-sequitur.* As Joseph LeDoux points out, most processes that are considered prototypical examples of cognition also occur without any conscious awareness, so the absence of conscious recognition cannot be used to infer anything about the cognitive or non-cognitive status of emotional appraisal (LeDoux, 1998).

By pointing to the powerful role of unconscious factors in preferences, Zajonc provided a useful reminder that introspection is of limited use when doing research in cognitive psychology. Unfortunately, this important reminder was somewhat obscured by the fact that Zajonc perpetuated the old mistake of equating cognition with consciousness.

*What Zajonc really meant*

As Paul Griffiths notes, the debate between Zajonc and his opponents about the relationship between emotion and cognition can sometimes seem like a mere semantic squabble, as if all it were about was the correct use of the term 'cognitive' (Griffiths, 1997: 25). However, there are more substantive issues at stake. No one denies that some kind of information processing must go on before an emotion can emerge in response to a given stimulus. The real question is whether or not this processing is of a kind sufficiently like the paradigm cases of unemotional thought-processes (such as deductive reasoning) to justify treating emotions as kinds of thought. Most emotion research in classical cognitive science has assumed a positive answer to this question. This is the assumption that Zajonc opposes.

As I noted in the previous section, mental processes are understood in classical cognitive science as bare computations. There is no provision here for distinguishing among different kinds of mental process, such as cognition and emotion. Zajonc is best construed as pointing out that the classical models of emotion do not provide a way of distinguishing emotion from cognition. If emotions are representational, what distinguishes them from thoughts, which are also representational? What Zajonc really meant was that we need to supplement the classical approach with new theoretical resources if we wish to do justice to emotion as a separate category of mental process.

At the time Zajonc wrote his paper, in 1980, the non-classical forms of cognitive science described in this thesis barely existed. In the years since then, the emotions have received much more attention from those working with non-classical approaches than they received from the pioneers of the classical approach. This suggests an intriguing possibility. Perhaps the non-classical approaches provide just the new conceptual resources needed for understanding the emotions that Zajonc was calling for in 1980. By providing new criteria for distinguishing between different *kinds* of computation, the non-classical approaches can help cognitive science find a new, non-Humean way of explicating the difference between cognition and emotion.

*The propositional attitude theory of emotion*

At the same time as cognitive scientists were developing appraisal theory, a group of analytic philosophers were working along similar lines. Like the appraisal theorists, these philosophers argued that emotions could be analysable purely as particular kinds of mental representation or thought. Specifically, it was proposed that emotions might simply be kinds of *judgement*. A concise statement of this view has been put forward by Richard Solomon:

> What is an emotion? An emotion is a judgement ...[For example,] my embarrassment is my judgement to the effect that I am in an exceedingly awkward situation... an emotion is an evaluative (or a normative) judgement.
>
> (Solomon, 1977: 185, emphasis in original)

Now, most psychologists and philosophers accept that some emotions, at least, *imply* certain judgements. It is hard to imagine how someone could feel guilty, for example, unless they judge, perhaps unconsciously, that they have done wrong. Solomon's claim, however, is much stronger. He claims that an emotion *is no more than* its constituent judgements. Emotions, on this account, are entirely reducible to (kinds of) thought.

This approach to emotion became known, among analytic philosophers, as the 'cognitive' theory of emotion. This name is misleading, since it suggests a concern with the findings of cognitive psychology, when in fact there was, at that time, virtually no contact between the philosophical and psychological investigations into emotion. I will therefore follow Paul Griffiths in referring to this philosophical tradition as the 'propositional attitude theory of emotion' (Griffiths, 1997: 2).

Although there was no real dialogue between the cognitive psychologists who developed appraisal theory and the philosophers who pioneered the propositional attitude theory of emotion, the two theories are remarkably similar. Both reject Hume's claim about the non-representational nature of emotion, and both attempt to reduce emotions to particular kinds of thought.

*Criticisms of the propositional attitude theory*

One of the most articulate philosophical critics of the propositional attitude theory of emotion is David Pugmire. In his book, *Rediscovering Emotion*, he argues persuasively that emotions cannot be reduced to judgements. His most powerful arguments turn on cases in which a person experiences an emotion of which the alleged constituent judgements are at odds with the person's explicit beliefs. Pugmire dubs such emotions 'irrational', and provides a number of examples. One example concerns a man who does not believe in ghosts and yet is overtaken by fear as he enters a deserted house. Another is that of a bereaved widow who finds herself angry at her husband for having 'left' her, even though she knows full well that his death was an accident. Such emotions are not reducible to judgements, Pugmire claims, since the judgements they might be taken to consist of conflict with the person's avowed beliefs.

Proponents of the propositional attitude theory of emotion have tried to deal with irrational emotions in a variety of ways. One is to say that the object of the emotion is represented less definitely in the person's mind; the man entering the deserted house does not believe it is haunted, since he does not believe in ghosts, but he does think that there is a danger of some unspecified kind. Another is to say that the belief is not certain; the man does not believe firmly in ghosts, but nor is he certain that they do not exist, and he is better off safe than sorry. Pugmire argues that not all irrational emotions can be explained in these ways. Sometimes a person's beliefs explicitly preclude the judgements alleged to be constitutive of the emotion, yet the emotion is still felt. The man explicitly states that he does not consider the house to be in the least dangerous, and yet he is still afraid when he enters it. So, Pugmire concludes, the propositional theory is in trouble.

Pugmire's objections to the propositional attitude theory fail, however, and for the same reasons as Zajonc's criticisms of appraisal theory: thoughts and judgements can be *unconscious*. Neither Zajonc's experiment with the mere-exposure effect, nor the cases of irrational emotion cited by Pugmire, rule out the possibility that emotions are reducible to *unconscious* thoughts. Like the appraisal theorist, the proponent of the propositional attitude theory can answer these objections by appealing to deeper cognitive resources offered by unconscious beliefs. When the man who is afraid of the deserted house tells us that he believes it is safe, there is no obvious reason why we should take his statement at face value. Perhaps he is simply *unaware* of his true beliefs. Or, perhaps he is only aware of *some* of his beliefs. If we take a domain-specific view of the mind, there could be various mutually inconsistent beliefs held by different mental parts, and it may be senseless to ask which of these beliefs is more truly *his*.

Pugmire is aware of the first of these responses, but not the second. When he points out that there are problems with tracing unconscious beliefs, then, he fails to see that this might be because we are assuming that the agent must have a single set of mutually consistent beliefs. Pugmire himself makes this dubious assumption in speaking of *irrational* emotions in the first place. He also assumes that, in cases of conflict between reported belief and a hypothetical unconscious belief, we must decide which of the two is the agent's *real* belief.

According to Pugmire, the claim that unconscious judgements must always have been at work in irrational emotions 'would rank as dogmatic, as an *ad hoc* hypothesis on the part of a psychoanalytically minded cognitivism' (Pugmire, 1998: 27). But this goes too far. Beliefs, or at least *representations*, are the fundamental explanatory tool of classical cognitive science. But cognitive scientists are not simply being dogmatic when they claim that that mental processes are thoroughly representational. There are good reasons that can be adduced to support this claim, some of which were outlined in the previous section. Among these reasons is the fact that we already have a good theoretical account of how representations can be processed by such obviously material entities as electronic computers, and thus a possible solution to the mind-body problem. In rejecting the representational theory of mind, Pugmire puts the mind-body problem back in the realm of mystery and thus leaves open the door to dualism.

Pugmire's main motivation for rejecting the representational theory of mind is his desire to save the Humean distinction between reason and the passions. He seems to think that, unless we can save this Humean thesis, we will have to give up the ancient distinction between cognition and emotion altogether. This, however, is clearly a *non-sequitur*. Just because we reject Hume's account of the distinction between reason and emotion does not mean that we have to give up the distinction altogether. There may be other ways of explicating this distinction, ways that do not appeal to a non-representational view of emotion. In the following chapters, I argue that the non-classical variants of cognitive science provide the resources that allow us to explicate the cognition-emotion distinction without giving up the representational theory of mind (and thus without giving up CTM).

# *Chapter Three:*

# *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. The 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 classial 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] 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

(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

advantage that favours one viable design over another, but on the putative *inviability* of one kind of design.

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 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

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

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 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 embarrasment 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 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, 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.

[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.

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 to 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 than 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 ammend 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

[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.

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 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,

[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.

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 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

[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). 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.

> 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 a 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 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.

# *Chapter Four:*

# *Situated cognitive science and emotion*

> *In contemplating your true self, don't include the body which surrounds you and the limbs attached to it. They are like tools, the only difference being that they grow from the body.*
>
> Marcus Aurelius, *The Meditations*

A growing number of cognitive scientists today argue for the need to take a 'situated' approach to the study of the mind. In this chapter, I outline the main principles of this approach, and show how they can be applied to the interruption theory of emotion.

## *4.1. Situated cognitive science*

When the term 'evolutionary' is used to describe an approach to the study of the mind, one immediately gets a pretty good idea what is involved. The same cannot be said of the term 'situated'. The label is due to Lucy Suchman, whose 1987 book, *Plans and Situated Action*, has been one of the most influential works in shaping this approach (Suchman, 1987). The core idea in Suchman's book, as I see it, is that minds are 'leaky', although this way of putting things is due to Andy Clark rather than Suchman herself (Clark, 1997).

### *The leaky mind*

In section 2.1, I identified internalism as one of the assumptions of classical cognitive science. I also argued that the best way to construe this term was as the view that the boundary of the mind is to be *identified* with some physical feature of the organism. This is the assumption that the proponents of the situated approach dispute. The input-output boundary is, according to them, a moveable feast. In other words, the question of where the input-output boundary is to be physically located will depend very much on the particular context of enquiry. The boundary of the *mind* is not to be simply *identified* with *any* physical boundary, whether that between the brain and the rest of the body or that between the body and the rest of the world. Some mental processes may well supervene entirely on the brain, or even on just one part of the brain. Others supervene on the brain plus part of the body. Others supervene on the brain, body, and parts of the world. This is what I mean by saying that minds are 'leaky'.

Why should we take the idea of the leaky mind seriously? Without attempting an exhaustive survey, I will summarise some of the arguments here. Since the most 'watertight' view of the mind locates it entirely in the brain, I begin by looking at cases in which the mind leaks out into other parts of the body. I then go on to examine cases in which the mind leaks even further afield, spilling out of the body into the world around it. If we wish to distinguish between these two cases of leakage, we could refer to the first as examples of the 'embodied' mind, and the second as cases of the 'embedded' mind.

### *The embodied mind*

The idea that the mind can be *embodied* is more than just the claim that natural minds are always found in creatures with bodies. The latter claim is just STM. To say that minds can be *embodied* is to say that mental processes do not always supervene simply on the brain; sometimes, they supervene on the brain *plus some other part of the body.*

One kind of argument for the idea that minds can be embodied in this sense appeals to cases in which changes to the body cause changes to the mind. For some of those afflicted by Möbius syndrome, for example, the inability to move any of the muscles of facial expression leads the person to experience significantly less affect than others (Cole, 1998). In such cases, it is not plausible to think of the body as simply a medium through which the mind expresses itself; the body affects the mind so powerfully that it seems more than just a figure of speech to say that it is part of the mind. Daniel Dennett makes a similar point in his book, *Kinds of Minds*, where he argues that the mind is 'distributed' throughout the body. In Dennett's words, 'my body contains as much of *me*, the values and talents and memories and dispositions that make me who I am, as my nervous system does' (Dennett, 1996: 77, emphasis in original)

Those who are tempted to dismiss such cases on the grounds that they do not deal with properly *cognitive* phenomena can be presented with other cases. People often use their fingers to keep a tally of something instead of storing the number in their heads. Less consciously, there is no need for the brain to calculate the force needed to sit up straight too exactly, since a certain amount of leeway is catered for by the purely physical properties of the body. Andy Clark takes these kinds of example to show that other organs besides the brain are integral to the processing loops that result in intelligent action. In other words, the brain can ease its computational burden by 'offloading' some information processing onto other parts of the body (Clark, 1997).

More generally, George Lakoff and Mark Johnson claim, in *Philosophy in the Flesh*, that the very structure of our thoughts depends on the particular kind of body we have. For example, 'the fact that we have muscles and use them to apply force in certain ways leads to the structure of our system of causal concepts' (Lakoff and Johnson, 1999: 19). According to Lakoff and Johnson, if we had different kinds of bodies, we would have different kinds of concepts. This thesis could be tested empirically. We might, for example, be able to design an experiment to test whether people with only one arm could conceive of 'weighing' and 'balancing' as easily as those with two.

### *The embedded mind*

According to the leaky mind hypothesis, the mind is not only capable of spilling out of the brain into the rest of the body, but also of flowing out of the body into the rest of the world. In addition to being embodied, minds can be *embedded.*

One set of examples to support this view involve 'cognitive artefacts' – objects made for the purpose of aiding cognition, such as maps, notepads and computers. Such objects not only lighten the processing load of the thinker (by allowing, for example, external memory storage), but also make possible certain tasks that would otherwise be practically impossible. Nicholas Humphrey puts the point well when he writes that a man with a pair of scissors 'is not just handier, he is in effect brainier – because he can now exploit his brain power in new ways' (Humphrey, 1997: 100).

Daniel Dennett and Andy Clark have both argued that the most important among the various mind tools at our disposal is language. The idea that external linguistic tools might alter and inform an individual's intrinsic mode of information-processing goes back at least to the Soviet psychologist, Lev Vygotsky, but for many years this possibility was largely ignored by many Western cognitive scientists, who followed Chomsky in adopting a thoroughly internalist approach to the study of language. Dennett and Clark have helped to redress this imbalance by showing how 'language is not just a medium *in which* we think', but in fact 'actually *does* some of the thinking with us and for us' (Humphrey, 1997: 101, emphasis in original). In Clark's terms, language is the 'ultimate (cognitive) artefact'; it provides the human brain with its most extensive and powerful 'external scaffolding'. Much of what humans take to be their own mental capacities, argues Clark, rely crucially on the external scaffolding provided by language (Clark, 1997).

The mind can leak out even further into the rest of the world, to include other people as well as cognitive artefacts. Edwin Hutchins proposes that the cockpit of an aeroplane be regarded as a cognitive system in which cognition is *distributed* over the pilot, co-pilot, and navigator, as well as over the various objects they use (Hutchins, 1995). There are, he claims, certain cognitive processes which simply cannot be located in one particular part of this system. Some of the complex capacities that we identify as mind and intellect may be much more like the systemic properties of the cockpit of an aircraft than intrinsic capacities of the bare biological brain.

### *The internalist reply*

The internalist can respond to all the supposed cases of embodied and embedded minds with a single, simple move; he can simply stick to his guns, and insist that what is going on in such cases is simply a case of a brain interacting with its environment. If we want to tell a richer story about how this two-way interaction goes on, all well and good, but the mind still supervenes exclusively on the brain.

The proponent of the situated approach cannot deny this move on purely logical grounds. He must concede that, for all the cases of leakage described above, it is always logically possible to redescribe them in such a way as to shift the input-output boundary back onto the skin of the organism, or further back still, onto the brain-body boundary. His reasons for rejecting such a move must, therefore, be of a less abstract kind than pure logic. This, however, is nothing new. Pierre Duhem and W. V. O. Quine showed long ago that scientific questions can never be decided on logic alone, nor on facts alone. Most philosophers of science today accept that our choice of alternative theories must be guided by other more pragmatic grounds too.

Two such pragmatic criteria present themselves as relevant to resolving the debate between internalism and the leaky mind hypothesis. The first is explanatory flexibility, and the second is empirical fruitfulness.

By explanatory flexibility, I mean that our explanatory tools must be able to accommodate themselves to different kinds of situation, rather than forcing the facts to accommodate to the theory. I recognise that this is a hopelessly sloppy way of putting things, but I take it that the reader has some idea of what I mean. The metaphors of the shoe-horn and the Procrustean bed are often used to describe inflexible explanatory frameworks. Of course, we don't want our framework to be *too* flexible, so that it can accommodate itself to any fact at all, for otherwise it will become vacuous. But nor do we want it to be so inflexible that every single situation receives the same pat response. That seems like dogmatism rather than science.

The internalist response seems to have all the hallmarks of dogmatism. No example of embodied or embedded minds, it seems, will be sufficient to persuade him that the input-output boundary is flexible. Yet he can give no principled reason for taking the skin or any other purely physical thing to be so crucial to the concept of mind. In fact, this seems to fly in the face of the key idea of substrate neutrality, which is so central to cognitive science. The brain is only one element in the processing loops that result in intelligent action, and to single it out as somehow 'more vital' to these action loops than any other seems somewhat arbitrary (Clark, 1997). Richard Dawkins has levelled the same charge of arbitrariness at biologists who identify the phenotype with the physical organism (Dawkins, 1982). Indeed, Dawkins' concept of the 'extended phenotype' can be seen as the biological analogue of the idea of embeddedness.

### *Parsimony*

The second pragmatic criterion that is relevant to resolving the debate between internalism and the leaky mind hypothesis is empirical fruitfulness. By empirical fruitfulness I mean that we should prefer one theory over another when it leads, as a matter of fact, to more interesting technical achievements. In the context of the debate between internalism and the leaky mind hypothesis, these technical achievements are to be found in the domain of artificial intelligence. Thus, if we find that cognitive scientists who adopt the leaky mind hypothesis tend to design better computing machines than those of a more internalist bent, this will constitute good grounds for preferring the leaky mind hypothesis. By their works shall ye know them.

The relevant sense of 'better' in this case is not merely one of performance, but one of accuracy in modelling natural minds. This assumes that we have some independent criterion, other than performance, for judging which of two machines is a more accurate model of the human mind. I think we do have such a criterion; it is the criterion of parsimony.

In this context, the criterion of parsimony states that whenever two machines are capable of mimicking some aspect of human behaviour equally well, we should assume that the simpler machine is a better model of the human mind. The justification for using this criterion is not just that simpler models are easier to work with. Nor is it just that the criterion accords with Occam's razor. It is also that, other things being equal, natural selection will favour simpler minds over more complex ones. To get this extra justification for the criterion of parsimony, of course, we need ETM.

When we compare the machines designed by proponents of the situated approach with internalist-type machines, the criterion of parsimony often favours the former. For example, how might we construct a robot that can pick up disused cans from the office floor? If we took a typical internalist approach we might assume that it was necessary to give the robot a sophisticated internal map of the office, and a complex camera system to register its position. The camera could search for tin cans at long range, in which case we would also need some complex image-analysis software. Alternatively, we could give the computer a short-range can detector and then make it traverse the whole office according to some pre-arranged systematic route. Either way, the program is going to be quite complex. Let's call this the internalist robot.

Connell, however, took a rather different approach. Inspired by the idea of offloading as much computation as possible on the environment, Connell simply allowed his robot to take a random route and gave it a host of infrared proximity sensors to help it navigate along walls and through doorways. Whenever it happened to detect a can with its short-range metal detector, it stopped to pick it up. This robot performed its task perfectly well without the need for any internal map of the room, nor for any pre-arranged route to be stored in its memory. As Rodney Brooks argues, it is natural to construe the robot built by Connell as being far simpler than the internalist robot (Brooks, 1991). This is only one example, but in many other cases too, the machines designed by those adopting a situated approach have managed to achieve comparable results to more internalist-style machines despite being much simpler. Overall, then, when judged by the criterion of parsimony, the situated approach has led to the design of better artificial minds than the internalist approach.

### *Heuristics*

There is nothing about CTM that forces one to adopt an internalist approach. CTM is perfectly compatible with the leaky mind hypothesis. Nevertheless, the fact remains that most classical cognitive *did* adopt an internalist approach. It seems that, unless people are forced to think otherwise by adopting an explicitly situated approach, they tend to operate on internalist assumptions.

Likewise, there is nothing about the internalist approach that *forces* one to design robots in one way rather than another. There is nothing about adopting the internalist stance that obliges the roboticist to design what I have dubbed 'internalist-robots'. Still, those who have adopted an internalist approach have tended to ignore the possibility that computational burdens can be eased by simple mechanical solutions. Even though there is nothing about the internalist position that logically excludes this possibility, it is the case that, as a purely empirical matter, cognitive scientists of an internalist bent have regularly ignored it. The situated approach recommends itself to us then, along practical grounds too. It is a heuristic that can guide those working in artificial intelligence to design their machines in simpler and more elegant ways.

I have argued that the situated approach can shed new light on the nature of mind. In the next section, I argue that the same is true when it comes to understanding that particular subset of the mental we refer to as emotions.

## *4.2. Situated approaches to emotion*

In section 3.2, I argued that new light could be thrown on the emotions by defining them as interruption mechanisms. In this section, I argue that we can advance our understanding of the emotions still further by applying a situated approach to the interruption theory.

The key idea of the situated approach is that mental processes are leaky. They do not always supervene on the brain alone. Sometimes they are 'embodied', in the sense of supervening on the brain plus some other part(s) of the body. And sometimes they are 'embedded', in the sense of supervening on the extra somatic environment too. In this section I argue that this is particularly true of those mental processes we call emotions. In other words, interruption mechanisms are very often embodied and embedded.

### *The internalist view of emotion*

The most obvious roles for bodily organs other than the brain in emotion is to serve as sensors for gathering emotionally relevant data and to serve as media for communicating emotions to other cognitive agents. These roles were clearly recognised by the proponents of the classical approach to emotion, but they tended to focus exclusively on *linguistic* inputs and outputs. The paradigmatic model of such a purely linguistic system is, of course, the Turing test, which (in its classical form) limits all communication to text messages sent via a keyboard to an alphanumeric display. In such a set-up, emotions can only be expressed by linguistic means. However, most emotion researchers today accept that the 'affective bandwidth' of linguistic media is very narrow, notwithstanding the expressive power of written dialogue that we find in great scripts and screenplays. Emotions are primarily communicated not by words but by a whole host of nonlinguistic bodily signals, from facial expression and vocal intonation to gestures and posture. To make our computer models of emotion more realistic, therefore, we should give them the ability to use the visual, auditory and physiological signals of emotion. To this end, roboticists and others working in the field of affective computing are increasingly supplying their machines with a wide variety of peripherals in addition to the standard keyboard and monitor. Peripherals that could be such as cameras and microphones are used to help computers get better at recognising emotions in humans, while animated agent faces and affective voice synthesisers are used to model nonlinguistic forms of emotional expression (Picard, 1997).

All this, however, is perfectly compatible with an internalist approach to emotion. Giving computers more peripherals than a keyboard and a monitor can certainly improve their ability to recognise and express emotions, but such technical advances do not seem to pose a serious theoretical challenge to the internalist view of emotion. The internalist can simply treat all the extra peripherals as providing richer input to, and more expressive output from, the central cognitive system where the 'real' emotional processes go on. All discussions of the 'recognition' and 'expression' of emotion are perfectly consistent, and perhaps even reinforce, a view of emotions as essentially *brain*-processes. The body just supplies input to the brain or serves as a vehicle for output, but the interruption mechanisms themselves are implemented exclusively in the brain. The input-output boundary is thus firmly maintained between the brain and the rest of the body.

### *Embodied emotions*

Classical cognitive science is not necessarily committed to the internalist view of emotion, and the internalist view of emotion does not logically entail that bodily inputs are non-specific (i.e. that bodily inputs do not determine which emotional response is produced). However, in practice most classical cognitive scientists did, in fact, take an internalist view of emotion, *and* assumed that bodily inputs were non-specific. The most famous proponents of this view were Stanley Schachter and Jerome Singer. On the basis of a well-known experiment on the emotional effects of adrenaline (Schachter and Singer, 1962), Schachter argued that emotions arise from cognitive attributions about one's state of physiological arousal (Schachter, 1964). In other words, all the body does is to tell the brain that *some* kind of emotional response is called for, without specifying *which* kind. The bodily input is merely a quantitative signal which, when it exceeds some threshold, triggers a cognitive process of appraisal such as those described in section 1.2. According to Schachter, it is this cognitive process that determines which emotional process will be set in motion.

This view of the role of the body in emotion goes back to Walter Cannon, who proposed it as an alternative to an earlier view put forward by William James (Cannon, 1927). James had argued, at the end of the nineteenth century, that all conscious feelings associated with emotions are merely the perception of physiological changes (James, 1888). This implies that bodily inputs are specific to each type of emotion. On the basis of some experimental work, Cannon argued that James was wrong, and that bodily inputs did not differentiate the various emotions from each other. Cannon's arguments won the day, and held sway for the rest of the twentieth century, seemingly confirmed by the famous 1962 study by Schachter and Singer. More recent experimental work, however, has led some researchers to call for a return to a more Jamesian view of emotion. Paul Ekman, for example, has published a study purporting to show that there are, in fact, distinctive ANS signatures for some of the so-called 'basic emotions', such as fear and anger (Ekman, Levenson et al., 1983). There are empirical grounds, then, for reviving the debate about the role of bodily inputs in emotional processing. It can no longer be assumed, as it was for much of the twentieth century, that bodily inputs are not specific to certain emotions. If Ekman's data can be replicated, we must broaden our view of where emotional computation takes place; it may not be the brain alone that instantiates this computation, but the brain plus other bits of the body.

The idea here is that the brain may offload some of its computational burden in emotional processes by storing some information about its current or recent emotional states in the body rather than storing it all in the brain itself. In that case, the emotional system consists not just of some neural structure but of the relevant brain structure plus some extra-neural bodily state. Once again, it would be open to the internalist to redescribe this in terms of an emotional brain mechanism monitoring input from the body as well as from the rest of the world. However, but since the interaction between brain and body is two-way, it allows for complex feedback loops that cross the boundaries between brain and body, so the emotional system is better seen as spread out over the whole agent. Andy Clark and Michael Wheeler have dubbed this effect 'causal spread'.

Thinking along these lines, Dolores Cañamero has written a computer program that simulates some of the physiological aspects of human emotions. The program runs in a two-dimensional virtual world inhabited by virtual creatures called 'Abbots' and 'Enemies'. Enemies do not have emotions, but Abbots have six basic emotions, each of which can be triggered by both external events and by internal physiological changes. The Abbots' virtual physiology includes synthetic hormones such as endorphins which, as in humans, can trigger a state of happiness or reduce the perception of pain (Canamero, 1997).

### *Embedded emotions*

Perhaps the claim that emotions are embodied is not that surprising. After all, the idea that emotions involve the body as well as the brain is probably more intuitive than similar claims about thoughts. The idea that emotions are embedded, however, is certainly not so intuitive. The claim that emotional processes supervene on bits of the *external world* as well as bits of the organism is so much of a challenge to the traditional internalist view of emotion that it needs to be argued for in more detail.

To illustrate what is meant by the claim that emotions are embedded, I will discuss a number of external devices that humans use to interrupt ongoing activity. Specifically, I will discuss two kinds of such external device: emotional artefacts, and emotional institutions. This is not meant to be an exhaustive taxonomy.

#### *Emotional artefacts.*

In the previous section, I mentioned a number of 'cognitive artefacts' such as maps and calculators. People use these devices to ease the burden of cognitive activity by partially offloading it onto the environment. In this section, I want to argue that humans also use a number of 'emotional artefacts'. These are devices that ease the burden of emotional (interruptive) activity by offloading it in a similar way.

#### *(i) Clocks*

Clocks may not appear to have much to do with emotion according to our pretheoretic understanding of the term, but if we accept the definition of emotion given in the previous chapter, we may decide we have to revise this initial impression. Clocks clearly function as interruption mechanisms. They may have other functions too, but it can scarcely be doubted that one of the main purposes for which clocks are constructed is to enable us to set time limits for various activities. Alarm clocks are designed to interrupt sleep. Timers are used to interrupt cooking, teaching, and many other activities. True, we may ignore them, but then we may sometimes decide to override our *internal* interruption mechanisms too, such as when a soldier persists in marching towards the enemy despite the feeling of fear that urges him to run away. If we combine the interruption theory of emotion with the leaky mind hypothesis, clocks may be regarded emotional artefacts.

It should hardly be necessary to repeat here that the normal internalist response is always possible. The internalist can simply argue that clocks provide external input into the real emotional system, which supervenes entirely on the brain or some part of it. It should be clear by now how the proponent of the situated approach to emotion can deal with such a response. The behaviour-generating system of interest here comprises the brain, the body and the clock. It is spread out over the agent and his environment. To insist that we must divide this system up along traditional lines by identifying the input-output boundary with the skin of the organism is to give up on the richer explanatory perspective offered by the situated approach. The purely logical possibility of re-analysing the system in internalist terms is not really all that exciting.

### *(ii) Fire alarms and burglar alarms*

Fire alarms and burglar alarms are also clear examples of interruption mechanisms. In particular, they partially offload the computational burden of the fear mechanism onto the outside world. It might be the case that, other things being equal, those who have such alarms installed in their houses tend to show fewer neural or physiological effects of fear in the relevant contexts than those without such alarms. I know of no evidence to support this hypothesis, but if it were true, it would constitute strong empirical support for the situated approach to emotion, for then we would have demonstrated very tight feedback effects between the neural/physiological aspects of the fear system and the external/mechanical aspects. Alternatively, if it could be shown that those with fire or burglar alarms show more neural or physiological correlates of fear in the relevant contexts than those without such alarms, we would also have evidence for the kind of feedback effects that favour a situated approach to emotion. The absence of feedback effects of either kind would not demolish the situated approach to emotion, but it would raise questions about its explanatory power.

### *(iii) Cruise-control*

Some cars are fitted with cruise-control. This device allows the driver to set a maximum speed beyond which the vehicle cannot accelerate. If the driver then tries to drive faster than the pre-set speed limit, the mechanism cuts in and temporarily disables the throttle. This device is clearly an interruption mechanism. Like the fire alarm and the burglar alarm, it eases the computational burden of fear by partially offloading it onto the car. The driver no longer has to worry about going too fast.

## *Emotional institutions*

The emotional artefacts described so far all offload fairly simple emotions like fear. What about other, more social emotions, of the kind discussed by Robert Frank? Can these too be partially offloaded onto the external environment?

As I have already noted in the previous chapter, Frank argues that many social emotions such as love and guilt function as commitment devices. I have already argued that commitment devices are just a special case of interruption mechanism, so I will not repeat the argument now. The important point to note here is that commitment devices can just as easily be offloaded onto the external world as other kinds of interruption mechanism. However, as befits their more social nature, these kinds of emotion are more typically offloaded onto the social environment than onto the physical one. They are offloaded, in other words, not onto physical artefacts, but onto social institutions. Such institutions therefore deserve to be called 'emotional'.

### *(i) Marriage*

Marriage is the emotional institution *par excellence*. Its function is to offload some of the computation involve in romantic love onto the social environment. According to Frank's evolutionary analysis, the function of romantic love is to ensure that you can form a stable pair-bond for long enough to raise at least one child. Since this function cannot be fulfilled without a multitude of conditions, romantic love may in fact designate a whole suite of emotions, rather than just one. For example, we might distinguish romantic fidelity from romantic generosity. These are not usually regarded as distinct emotions, but there are many reasons why common sense views of love might be more misleading than many other aspects of folk psychology, so we should not worry too much if the cognitive science of love leads to counter-intuitive findings.

If this view of love is correct, we need not be surprised if we find that marriage is a way of offloading *some* aspects of romantic love but not all. We can take each aspect of romantic love separately, and ask, for each one, whether or not it is plausible to see marriage as a way of offloading it.

As far as romantic fidelity goes, marriage seems to fit the bill as a means for offloading. Once two people have declared their vows in public, they do not always need to interrupt themselves when faced with an offer of extra-marital sex. If there are friends and relatives around, they will probably exert social pressure to make sure the vows are kept.

Marriage is also a way of offloading romantic generosity. The common practice of exchanging gifts at wedding anniversaries means that there is, in effect, an external timing mechanism that repeatedly interrupts the day-to-day routine of married life and prompts the giving of gifts. Here again, the scrutiny of friends and relatives can help to enforce this interruption mechanism.

### *(ii) Confession*

By means of the institution of confession, many Catholics partly offload guilt. The rule forbidding one to take communion while in a state of sin, coupled with the expectation that one will take communion fairly regularly, means that there is in effect a kind of external alarm clock that repeatedly forces one to examine one's conscience. It is easy to see how this institution is a way of offloading the computational load of constantly checking to see if you have kept your promises. When you have not, you must tell a priest, who then exerts social pressure on you to make ammends.

### *Emotional groups*

From what has been said so far, one might easily gain the impression that emotions can only be offloaded onto the environment by creatures like ourselves, who are sophisticated enough to dream up ways of doing so. This, however, is not the case. For those animals with less sophisticated minds, natural selection can do the design work instead. Merekats did not sit down and work out their system of taking turns to do sentry-duty; the system evolved by natural selection, without any conscious reasoning on their part. Yet this too is plausibly described as an embedded emotion. Those merekats that are not doing sentry duty have partially offloaded the computational burden involved in detecting predators onto part of their social environment. They have, that is, partially offloaded the emotion of fear onto another merekat.

Humans, too, offload some emotions onto the social environment without the need for social institutions. The power of crowds to amplify the emotions of each individual member is a case in point. Social psychologists have long remarked on the distinctive nature of emotional behaviour in groups, from the contagious nature of panic to the lynch-mob mentality in which otherwise peaceful people can sometimes be caught up. These too are plausibly described as cases of embedded emotion, when the feedback loops between brain, body and social environment mean that the whole is greater than the mere sum of its parts.

### *Purely cerebral emotions*

Adopting a situated approach to emotion does not imply that all emotions are embodied and embedded. Quite the contrary, in fact. The whole point of the situated approach is not that the input-output boundary is located at some point other than the brain-body boundary; it is that the boundary is moveable. It is not enough to give examples, then, of how emotions can be embodied and embedded; we must also find plausible examples of purely cerebral emotions. Otherwise, the mind would not be *leaky*; it would be permanently *overflowing*.

This is not a problem. There are plenty of examples of emotions where the behaviour-generating system of interest is the bare biological brain. If there were not, the internalist view of emotion would probably not have been so initially tempting. The startle reaction that makes a lone squirrel interrupt its foraging behaviour when it perceives a sudden noise may be an example of such a purely neural emotion. It need not be, however. If the squirrel's extra-cerebral body chemistry has been sufficiently altered by previous startle reactions to warrant being regarded as a repository of emotional memory, then this would be a case of an embodied emotional process. Everything depends on the details of the particular situation. The virtue of the situated approach is that it calls our attention to such details.

### *Situated emotions and artificial intelligence*

If the situated approach to emotion is to count as part of cognitive science, it must lead us to design better artificial emotion systems. I believe it can.

In particular it suggests that those working in affective computing do not need to build interruption mechanisms entirely *inside* the machine. When constructing emotional robots, designers can be sensitive to the possibility of exploiting features of the external environment to offload some of the computational burden.

However, the main lesson for affective computing to learn from the situated approach to emotion may be precisely the opposite. If computers seem so unemotional today, this is because they have, in a sense, already offloaded *all* their emotions onto their human environment. Computers do not *need* interruption mechanisms at the moment, because humans take care of all their survival and reproductive requirements. If more interruption mechanisms were installed in desktop computers, so that they shouldered more of the computational burden of caring for themselves, humans would find that their own burden became lighter. Computers would become easier to use.

Things began to move in this direction with the invention of screen-savers, which can thus be seen as the first *internal* emotions in desktop computers. As the example of the screen-saver shows, giving computers emotions can benefit their human users – in this case by allowing monitors to live longer and thus allowing human owners to save money by replacing them less often. Building in other kinds of interruption mechanism may make computers even more user-friendly. For example, a computer that was designed to interrupt activity momentarily when it detected that its user was becoming too tired might help to avoid faulty data being stored in its memory. Computers that become bored when they are not receiving input of a sufficiently novel kind (as in the ACRES program described in section 3.2) might help their users to avoid wasting time. And so on.

These points become even more relevant when the computers in question are designed for entertainment purposes rather than more utilitarian ones. The simulated agents in gaming software will be more realistic, and therefore more entertaining, when they have more sophisticated emotional capacities. This suggests that developments in affective computing over the next few decades may be driven more by the demands of computer gaming than by the demands of industry or commerce.

# *Chapter Five:*

# *Dynamical cognitive science and emotion*

*What is passionate in us rises and falls, leaps or creeps, and slowly paces. Now it becomes urgent, now hesitant, now stirred more feebly, now more strongly.*

Johann Gottfried Herder

In the past decade, there have been a number of calls for cognitive scientists to adopt a more 'dynamical' approach to the study of the mind. In this chapter, I ask what such calls actually mean. In the second section, I examine how the dynamical approach can help to refine the interruption theory of emotion.

## *5.1. Dynamical cognitive science*

The machines designed by cognitive scientists in the period 1950-1980 were almost all discrete-state machines. In such machines, the transitions between one state and another are like sudden jumps or clicks. There are no intermediate positions between one state and another.

Many of the first generation of cognitive scientists assumed, not only that this was the best way to design *artificial* minds, but that discrete-state machines would also be the best way to model *natural* minds. This assumption, however, is logically independent of both CTM and the design-based approach. The assumption of discreteness should not, therefore, be seen as an essential part of cognitive science *per se*, but simply as a characteristic feature of *classical* cognitive science. Its widespread acceptance among the first generation of cognitive scientists was a historical accident, not a logical necessity. This only became clear, however, when various sections of the cognitive science community began to propose other ways of building artificial minds.

The first cognitive scientists to do this were the connectionists. Rather than using a few heterogeneous components, each of which makes a specific isolable contribution to overall performance, connectionist models (often called nerual networks) employ a large number of homogeneous components. Whereas discrete-state machines usually process information serially, connectionist models are massively parallel (hence the term PDP, or 'Parallel Distributed Processing', that is often applied to such models).

The connectionists, however, were not always explicit about their commitments to continuous mathematics. Very often, they simply got on with the business of designing neural networks, and left the theoretical issues to philosophers. It was not until the emergence, in the 1990s, of proposals for a distinctively dynamical approach to cognitive science that the issue of continuity was really foregrounded.

Dynamical systems are not all continuous. There are discrete dynamical systems as well. But most proponents of the dynamical approach to cognitive science have tended to concentrate on continuous systems. It is therefore plausible to argue that the main contribution of the dynamical approach has been the challenge that it poses to the classical assumption of discreteness.

*The dynamical hypothesis of cognition.*

Dynamical cognitive science has been defined by its adherence to the 'dynamical hypothesis of cognition', which simply states that 'cognitive agents are dynamical systems' (van Gelder, 1998: 615). The term 'dynamical system' may be understood in various ways, but proponents of dynamical cognition usually construe it in very broad terms, as designating any system in which a variable, *x*, changes continuously over time. The 'system' in question may be (Smith, 1998: 4-5):

(i) *a real-world system*, such as the planets in motion or – in dynamical cognition – a cognitive agent, or

(ii) *a mathematical system*, which, in dynamical systems theory, may be (a) *a set of dynamical equations,* which perhaps aims to capture the behaviour of some real-world system, or (b) *an abstract mathematical structure*, such as a set of trajectories in a phase space which is characterised by a set of dynamical equations.

This multiple use of the term 'system' by the proponents of dynamical cognition is potentially harmful, as it might lead to ontological questions about the nature of things being confused with epistemological and methodological questions about the mathematical tools we use to describe and analyse things. To prevent such confusion, Tim van Gelder argues that we should distinguish two components of the dynamical hypothesis of cognition. On the one hand, 'the *nature* hypothesis is a claim about the nature of cognitive agents themselves: it specifies what they *are* (i.e. dynamical systems)' (van Gelder, 1998: 619, emphasis in original). The *knowledge* hypothesis, on the other hand, is a claim about cognitive science: namely, that we can and should understand cognition by using the resources of dynamical systems theory such as dynamical equations and geometrical modelling. Before going on to give some examples of research in dynamical cognition, I will briefly discuss the nature hypothesis and the knowledge hypothesis in turn.

*The nature hypothesis*

The nature hypothesis states that cognitive agents are dynamical systems (where 'system' is used in sense (i) described above). This purports to be an empirical claim about a set of real-world systems. However, things are not that simple. In particular, two points need to be made clear.

Firstly, it is rather misleading to claim that the statement 'cognitive agents are dynamical systems' amounts to a hypothesis about cognition, since this seems to imply that it makes some specific claim about cognitive agents. It sounds as if the property of being a dynamical system is being used to pick out cognitive agents from other kinds of physical entity. Yet this is not the case. *Any* physical system, from a single neuron to a galaxy, may be described as a dynamical system.[1] The concept of a dynamical system is no more specific than the rough idea of a computer as a system that transforms input into output. In fact, the two ideas are equivalent. To pick out those dynamical systems that are cognitive agents we need some further constraint, just as we did when we were trying to specify the notion of computation involved in CTM. We could, in fact, use the very same constraint that we used then – the notion of representation. We could, in other words, define cognitive agents as dynamical systems that represent other other dynamical systems.

On this account, dynamical cognitive science is thus no less 'computational' than other forms of cognitive science. Computers are just dynamical systems that represent other dynamical systems. When proponents of the dynamical approach, then claim that the view of the mind as a dynamical system 'is an entirely different image from the brain as a computer with stored contents or subroutines to be called up by a program' (Kelso, 1995: 1), they must clearly have in mind some much more specific notion of computation than the one I proposed in chapter one. They probably have in mind the classical view that defines computation as a process that terminates after a finite number of basic operations specified by an algorithm (in technical terms, an 'effective' process). This indeed, is Church's thesis, which stipulates that a process could only be *called* 'computable' if it is effective in this sense. Church's thesis has acquired such standing in computational theory that it

[1] To be more precise, any physical system can be construed as instantiating infinitely many dynamical systems. I come to this point shortly.

sometimes seems to be regarded as an empirical discovery, instead of the stipulative definition that it really is. Sterile disputes about whether or not the dynamical approach counts as 'computational' or not could be avoided if it were kept in mind that the term can have two meanings. According to Church's thesis, computers are just finite-state machines.[2] According to the broader definition I proposed in the first chapter, finite-state machines are just one kind of computer; you can have continuous computers too.

The second point that I want to make about the nature hypothesis is this: to say that a real-world system, such as a convecting fluid, is a dynamical system is actually to make a claim about instantiation, not about simple identity. A convecting fluid is not itself a set of variables, but rather a material entity whose behaviour can be *described* by a set of variables. Likewise, cognitive agents are not sets of variables, but they may behave in ways that are describable in such terms. Thus the nature hypothesis should be understood as claiming that cognitive agents *instantiate* dynamical systems (van Gelder, 1998: 619).

Furthermore, since any real-world system instantiates *numerous* dynamical systems, the nature hypothesis should not be construed as claiming that each cognitive agent is some *particular* dynamical system. Rather, it should be construed as claiming that each cognitive agent 'is' as many dynamical systems as are needed to describe all the different kinds of cognitive performance exhibited by the agent (van Gelder, 1998: 619).

The fact that claims about the dynamical nature of cognition are really just claims about the various dynamical systems that cognitive agents instantiate threatens to undermine van Gelder's distinction between the nature hypothesis and the knowledge hypothesis. Claims about instantiation are reducible to claims about the theoretical tools that are

[2] A finite state machine is a discrete state machine that consists of a *finite number* of discrete states, together with their state transition rules.

most appropriate to use in studying a given object. The nature hypothesis thus reduces to the knowledge hypothesis.

*The knowledge hypothesis*

The knowledge hypothesis of dynamical cognition is much easier to state clearly than the nature hypothesis. It is simply the claim that cognitive agents are better understood by appealing to the resources of dynamical systems theory. Dynamical systems theory is a branch of pure mathematics concerned with the properties of dynamical systems (where 'system' is used in sense (ii) described above). Typically, dynamical systems theory uses a set of linked differential equations to specify the evolution in time of some variable, *x*. In such systems, time is continuous. But there are other dynamical systems, specified in terms of difference equations, in which time is discrete. So continuous dynamical systems are really just a subset of the class of all dynamical systems. However, for the reasons given above, this chapter deals only with continuous dynamical systems. To re-cap: a theory of minds as *discrete* dynamical systems is already available in the classical approach. The novelty of the dynamical approach consists, therefore, in providing an alternative theory of minds as *continuous* dynamical systems.

Still it remains to be seen how much of an 'alternative' this theory really amounts to. Some of the proponents of the dynamical approach write as if cognitive scientists are faced with a stark choice between it and the classical approach (e.g. van Gelder, 1998). What they mean, I suppose, is that cognitive scientists much choose between discrete and continuous models of cognition; the dynamical approach is not necessarily committed to rejecting the other assumptions of the classical approach (domain-generality and internalism). Yet the choice between discrete and continuous models is far from being a black-or-white one.

For a start, the distinction between discrete and continuous systems is a mathematical one, and it is not clear how to apply it to real-world systems.

When a discrete model is criticised on the grounds that the real-world system is better represented by a continuous model, this can usually be reduced to one of the following two claims:

(i) The scale chosen by the discrete model is not sufficiently fine-grained: the dimension in question is usually (but not always) a *temporal* one.

(ii) The discrete model needlessly imputes too much internal structure to the real-world system.

However, since both of these criticisms can often be met by refining the discrete-state model, the real point at stake does not seem to be one continuity as such. I will now explain this point in more detail.

*The grain problem.*

Sometimes, I claim, the criticisms levelled against discrete models are reducible to the claim that the scale chosen by the discrete model is not sufficiently fine-grained. An example from the history of biology may serve to make this point clear.

In the first decades of the twentieth century, there was a rather silly feud between two ways of thinking about heredity. The Mendelians took a discrete approach; they argued that phenotypic characters were controlled by genes that were either present or absent (more precisely, only one particular *form* of a gene – one *allele* – could be present at any given chromosomal locus). The biometricians thought this was clearly at odds with the fact that some phenotypic characters can vary continuously; people are not, for example, just *tall* or *short.* Ronald Fisher showed that the disagreement could be solved if one assumed that such continuous phenotypic characters are polygenic – that is, controlled by more than one gene. If height is influenced by many genes of small effect, then it is clear how it can be approximately continuous.

The term 'approximately' is crucial. Whenever a discrete model differs from a continuous one, it can be made to approximate it more closely simply by giving it more discrete states. In other words, discrete models can always be made more realistic by choosing a more fine-grained scale of analysis. So when a discrete model is contrasted unfavourably with a continuous model, the charge often amounts to no more than the claim that the discrete model is not sufficiently fine-grained. And this criticism will only be valid if we can show that the scale chosen was not sufficiently fine-grained to meet the explanatory purpose for which the model was constructed. After all, the whole point of a model is not to be as fine-grained as the thing it is supposed to represent. It would be silly to criticise a map for not being as detailed as the terrain itself.

This point is borne out when one looks at the actual machines built by those who take a dynamical approach. These machines are usually just versions of connectionist networks. But connectionist networks are rarely built out of analogue components. They are usually simulated on discrete-state machines; the theoretically *continuous* activation level of each node in the network is, in practice, approximated by means of a fine-grained *discrete* series.

If connectionist networks can be implemented, to a good enough degree of approximation, on discrete-state machines, the reverse is also true. This, at least, seems to be what Fodor and Pylyshyn claim in their influential paper on connectionism and cognitive architecture (Fodor and Pylyshyn, 1988). They argue that a neural network architecture can implement the digital processes that they take to lie at the core of human cognition. If we combine these two points, we can imagine a machine that is digital at one level instantiating a system that, at a higher level of analysis, is continuous, and that this continuous system then instantiates a discrete-state machine at an even higher level. Thus the grain-problem does not just refer to the number of units on a particular scale; it also refers to the level of nature we are analysing (though perhaps this amounts to the same thing).

The dimension of discrete models which proponents of dynamical cognition most often focus on in their pleas for a finer grain of analysis is the temporal one. In other words, criticisms of discrete models often amount to no more than the charge that they have not been sufficiently sensitive to the details of timing. This may be a serious defect with some discrete models, but if so, it is not because they are *discrete*. It is just because they are not sufficiently aware that, for many cognitive processes, every millisecond counts. This problem can always be remedied by choosing smaller units for the temporal scale.

*Parsimony*

Whenever the criticisms levelled at discrete models cannot be plausibly construed as pleas for a finer-grain of analysis, they are usually reducible to the claim that the discrete models needlessly impute too much internal structure to the real-world system. To understand the connection between the idea of continuity and the idea of complexity, we need to make a brief digression into chaos theory.

Chaos theory is simply another name for the branch of dynamical systems theory that studies the properties of dynamical systems governed by nonlinear equations (Stewart, 1989: viii). An equation is linear if the sum of two solutions is itself a solution. The solution for a two-stone disturbance of a liquid surface, for example, is just the sum of the solutions for two one-stone disturbances, centred at appropriate points (Stewart, 1989: 72). Most classical equations, including those of classical dynamics, are linear. This is not true of the equations in chaos theory.

In chaotic systems, trajectories from nearby initial conditions can lead to outcomes that are not correlated. In other words, these systems exhibit sensitive dependence on initial conditions, a phenomenon that is sometimes known as 'the Butterfly Effect'.[3] Some popularisations convey the impression that this is the most original 'take-home message' of chaos theory (Stewart, 1989), but this is misleading to say the least. It is no news that small changes can have huge effects; Darwin once remarked that his voyage on the *Beagle*, which determined his whole career, had at one point depended on such trifles as the shape of his nose.[4] The simple idea that small changes can cause large, unpredictable effects has been around for much longer than chaotic dynamics (Smith, 1998: 1).

Chaos theory gives to a new twist to this old idea by showing that the very complex patterns formed by the trajectories in a chaotic phase-space can be produced by relatively simple equations.[5] A very complex series, which may even appear to be completely random to the untrained eye, can be produced by a deterministic formula.

The take-home lesson here is that when we observe some very complex behaviour, we should not assume that the system producing it necessarily has a complex internal structure. Now, it is probably fair to say that many of the first cognitive scientists tended to make exactly this assumption. Guided by the strategy of functional decomposition (see section 1.1), they tended to assume that all complex behaviour could only be generated by a

[3] The name comes from a story which is often used in the literature to illustrate the phenomenon of sensitive dependence on initial conditions. Here is Peter Smith's version of the story: 'A small blue butterfly, let's suppose, sits on a cherry tree in a remote province of China. As is the way of butterflies, while it sits it occasionally opens and closes its wings. It could have opened its wings twice just now; but in fact it moved them only once. And – because the weather system exhibits sensitive dependence – the miniscule difference in the resulting eddies of air around the butterfly eventually makes the difference between whether, two months later, a hurricane sweeps across southern England or harmlessly dies out over the Atlantic. Or so the story goes.' (Smith, 1998: 1)

[4] Robert Fitzroy, the captain of the *Beagle*, was a firm believer in physiognomy, according to which a person's character could be discovered by attending to the shape of their facial features. Darwin's nose seemed to Fitzroy to betray a lack of resolution, and thus almost cost Darwin his place on the ship (Bateson and Martin, 1999: 123-24).

[5] There seems to be some sleight of hand here. Unless we can specify some uniform measure of 'complexity' that applies equally to mathematical equations and to the patterns formed in a multi-dimensional phase-space, then it makes no sense to marvel at how wonderfully 'complex' instances of the latter can be produced by such 'simple' instances of the former.

set of heterogeneous components, like those in, say, a television or an electronic computing machine *circa* 1960. One of the surprising things about connectionist models was that they put this assumption in doubt. Some kinds of task, such as pattern recognition, could be achieved by very simple networks consisting of a few dozen homogenous components (simple nodes).

### *Hype*

If I am right, and most (perhaps *all*) criticisms levelled at discrete models can be reduced either to a plea for a finer grain of analysis or a call for greater parsimony, then the problem with such models is not their discreteness. Such criticisms could be met without making the model continuous. In neither case is the point at stake one of continuity as such.

Nor is it strictly necessary to appeal to the resources of dynamical systems theory to make these points. The point about choosing the appropriate grain of analysis can be made quite well without them. And, as the example of connectionism shows, it is not necessary to invoke the arcane terminology of chaos theory to make the point that complex behaviour can sometimes be generated by relatively simple systems. The connectionists did not often describe their models in terms of chaotic attractors, but they succeeded in challenging the assumption of internal complexity that seemed to underlie many classical models.

One is left with the impression that the real value of dynamical systems theory for cognitive science is rhetorical. The arcane mathematical terminology is being used not for any intrinsic value, but merely to make people think that there *is* something important being said here. Now, sometimes there is something important. Sometimes the discrete models do need to be finer-grained and more parsimonious. But these points would, I think, be made more persuasively if they were couched in more simple terms.

Kelso and other dynamicists are well aware of these suspicions, and try to distance themselves from the hype and the vague analogies that have characterised some of the recent writings on chaos by insisting on precise links between theory and data:

> I am not going to comment on the rhetoric surrounding the buzzword chaos and how it provides a more holistic view of human life, except to say, chaos of what? What are the relevant variables that are supposed to exhibit chaotic dynamics? What are the control parameters? And how do we find them in complex living systems where many variables can be measured, but not all are relevant?...What are the attractors? What does the bifurcation diagram look like? Are these concepts and mathematical tools even relevant? How does one establish them, even in a single case?...There has to be some connection between mathematical formulae and the phenomena we are trying to understand....Establishing a connection between theory and experiment is one of the canons of science that the 'chaos, chaos everywhere' crowd seems to ignore.
>
> (Kelso, 1995: 43-44)

Kelso rightly draws importance to the importance of empirical research, and his book includes some of the best examples of such work. One such example is the Haken-Kelso-Bunz (HKB) model of limb coordination. Kelso and colleagues studied coordination within and between limbs, and found that a whole range of different coordination patterns could be modelled with the same dynamical equation. The only variable required for this equation was the relative phase of the limbs in question. Relative phase refers to the relation between two oscillating components. Imagine that you are tapping the table in front of you with the forefinger of each hand. If the fingers hit the table simultaneously each beat, the relative phase is said to be 'inphase', while if the fingers hit the table alternately, the relative phase is 'antiphase'.

Kelso also found that a whole variety of limb coordination patterns, from the case of finger-wagging just described, to trotting and galloping in quadrapeds, could be modelled by a dynamical equation based on the derivative of the relative phase. This simple model predicted a wide of range of observed phenomena, including the small range of stable coordination patterns and the nonequilibrium phase transisitions that occurred when the system moved from one stable pattern to another (Kelso, 1995: 46-57, 74-87).

*The limited applicability of the dynamical approach*

The empirical work described by Kelso, such as the HKB model, goes some way towards pre-empting the charge of 'hype' But finger wagging and limb coordination are not exactly paradigms of 'mental' processes. Ture, movement is vital to all natural cognitive agents, but classical cognitive science did not completely disregard motor control. On the other hand, classical cognitive science also provided models for more paradigmatic mental processes like reasoning and problem-solving, but dynamical models for such things are practically non-existent. Even if we grant that there are natural advantages to modelling motor-control in continuous terms, it seems hard to believe that we could say the same about other mental processes.

Evolutionary cognitive science can suggest a very good reason for this. Natural minds evolved to guide adaptive behaviour; and it would almost never be useful to have cognitive or emotional systems do large numbers of iterations through dynamical states into order to achieve such behaviour.[6] Rhythmic locomotion seems to be the only case where coupled oscillators have a distinct advantage over discrete-state machines. No wonder, then, that dynamical models of 'cognition' tend to concentrate on things like limb-coordination.

[6] I owe this point to Geoffrey Miller (personal communication).

Even if we include connectionist networks in the class of continuous systems, the dynamical approach does not go much further. The only area in which connectionist networks seem to have a distinct advantage over classical discrete-state machines seems to be in the area of pattern recognition. Insofar as continuous models can be taken as a separate class from discrete models, then, their value may be restricted to 'low level' processes such as perception and motor-control.

### *Is the dynamical approach design-based?*

Proponents of the dynamical approach point out that connectionism is not the same thing as dynamics (van Gelder, 1998: 661). But this only weakens their claim to make a distinct contribution to cognitive science. The distinctive thing about cognitive science, I argued in chapter one, is its design-based methodology. If the dynamical approach cannot claim connectionist networks as paradigms of the continuous models they prefer, then it the dynamical approach is reduced to mere theory without any concrete proposals to show in the way of working machines. If the dynamical approach is to prove itself as a species of cognitive science, it must provide clear evidence of how it leads to new insights in artificial intelligence. So far, it has failed to do so. The various 'dynamical machines' all turn out to be old-fashioned connectionist networks, with perhaps the twist that several such networks are linked up in novel ways (see, for example, the robot described in Tani, 1999: 157).

Besides, even if we grant, for the sake of argument, that connectionist networks *are* distinctively dynamical machines, it is still not clear whether this will permit the dynamical approach to call itself design-based. The point of the design-based methodology is not simply that we build working machines, but that we understand *how* the machines work. With traditional digital machines, this is no problem. One can apply the strategy of functional decomposition quite easily to such machines. But connectionist networks are no so amenable to this explanatory strategy. There is no obvious way of assigning different functional systems such as 'memory' or 'executive' to a particular bit of the network. Rather, these functional systems tend to be 'distributed' across the whole network in a way that defies clear explanation. Connectionist networks thus seem to be much less 'transparent' in their workings than digital machines. This may be why people turn to mathematical equations to understand them; there is simply no other way. But describing a machine in terms of the mathematical function it computes is rather different to breaking it down into distinct subsystems. To the extent that the design-based approach of cognitive science requires functional decomposition, connectionism does not count as part of cognitive science.

## *5.2. Dynamical approaches to emotion*

In section 3.2, I proposed, on the basis of various evolutionary hypotheses, a definition of emotions as interruption mechanisms. In section 4.2, I argued that we could take this theory further by drawing on the insights of situated cognitive science. Can we take it further still by drawing on the insights of the dynamical approach?

These insights, I have argued, reduce to two points. To recap, if the dynamical approach offers anything distinctive, it is a reminder that we should be careful not get the grain of our analysis right, and not to assume that complex internal structure is necessary to generate complex behaviour. Let us see how we can use these insights to refine the interruption theory of emotion.

### *Are there enough states in the model?*

The classical models of emotion described in chapter two, such as the OCC model, represent emotions in discrete terms. In response to a given input, a description of an emotion is either generated or it is not. There are no intermediate states between the full presence of an emotion and its complete absence.

In humans and other animals, however, emotions are not such black-and-white affairs. They have quantitative characteristics as well as qualitative ones. At any one moment, an emotion may be present in varying degrees of intensity, and this intensity may wax or wane with varying rapidity. These quantitative aspects of emotion can be modelled in a discrete-state machine by adding more states. Rather than just assigning one bit to emotion (is it on or off?), we can assign several, representing degrees of emotional intensity. Or we can use a more obviously analogue system, such as a connectionist network.

Juan Velasquez of MIT has developed a connectionist model of emotion called 'Cathexis' (Velasquez, 1996). The network consists of three layers of nodes. In the input layer, the nodes represent the four kinds of emotional elicitor listed by Caroll Izard in her theory of emotion: neural, sensorimotor, motivational and cognitive (Izard, 1993). In the middle layer, each node represents an emotion such as joy or distress. In the output layer, each node represents a behaviour, such as smiling. The nodes in the middle layer are connected to each other as well as to the nodes in the other layers, so, for example, joy can inhibit distress and activate hope. Each emotion-node has a continuously variable level of activation, which represents the intensity of that emotion at a given point in time. Unlike the classical models, then, in which only one emotion may be present at a given time, all emotions are constantly activated in Cathexis, though at varying levels of intensity. The intensity of each emotion changes at regular intervals in accordance with an equation whose terms include the inputs from other nodes modified by inhibitory and excitatory gains, and a function that controls the temporal decay of the emotion. The new intensity is thus a function of its decayed previous value, the effects of its elicitors, and the influences of other emotions.

Cathexis manifests the quantitative features of emotion that the classical models leave out. Emotions are not simply present or absent, but are continuously present in varying degrees of intensity. Furthermore, the intensity of each emotion changes at differing rates. The equation that specifies how the intensity changes from one moment to the next is nonlinear, so chaotic dynamics may be observed. For example, since the equation includes parameters that specify a minimum activation threshold and a maximum saturation value, the possible intensity values for each emotion may be graphed as a sigmoidal curve in which the middle region is highly sensitive to initial conditions. The steepness of the sigmoid can be altered by changing the values of these parameters, which allows for different temperaments to be modelled. A steep sigmoid, for example, would reflect an emotionally labile temperament, while a gentler slope would represent a more phlegmatic character. Finally, the multiple interactions among the nodes allows for feedback loops with various nonlinear properties such as time-dependence (i.e. the effect of an elicitor on the intensity of an emotion depends on the time when the elicitor comes into play).

The interruption theory could learn from this model. I have already mentioned the possibility that the top-down effects of cognition on emotion might be modelled by allowing the top layer in the hierarchy (the cognitive layer) to have some control over the activation threshold of the lower layers (the interruption mechanisms). This is already to introduce the idea of variable intensity into our model. Cathexis suggests that we might extend this idea to interactions between the lower layers themselves. If the activation threshold of the lower layers was not influenced just by the top layer, but also by other lower layers, then feedback effects among the various layers could emerge. Increasing tiredness could, for example, lower the activation threshold for anger; this would model the tendency for tired people to lose their temper more easily than others.

Whether or not the addition of these quantitative features should be seen as inherently non-classical is dubious, however. They could be built into the interruption model without using a connectionist architecture, simply by using discrete-state machines with many more states to approximate the continuously variable intensity of emotions. Even if we did build some of the layers along connectionist lines, we would not need to build all the layers along such lines; we could use a hybrid architecture. And even if the machine were entirely composed of connectionist networks, it would still have features that could be described in discrete terms. Cathexis, for example, though touted as a thoroughly dynamical model of emotion, has some discrete aspects. Thus, while the intensity of each emotion appears to vary continuously, this is only because the discrete series of intensity values has many members. Furthermore, the fact that each emotion node only excites other nodes when its intensity surpasses a given threshold bestows a digital character on excitation.

### *Mood*

Building variable activation thresholds into the interruption theory may provide us with a good way of understanding what moods are. The features that are usually used to distinguish between emotions and moods are all phenomenological rather than mechanistic. Moods are usually said to build up and die away more slowly than emotions, to last longer than emotions, and to constrain attention less forcefully than emotions . If this is all there is to the emotion-mood distinction, we need not regard moods as separate mechanisms from emotions but merely as a set of dimensions along which various features of the emotion mechanisms can vary. An irritable mood, for example, could be modelled by simultaneously altering various parameters of the anger mechanism: for example, we might lowering the activation threshold, decreasing onset time, and increasing offset time. Let us call this the 'parameter theory of mood'.

This way of understanding mood would actually correspond quite well with the explication of mood offered by Vincent Nowlis. Nowlis suggested that moods were second-order dispositions (Nowlis, 1963). In other words, while emotions are dispositions to act in particular ways, moods are dispositions to have certain emotions. The parameter theory of mood provides a concrete way of understanding this rather abstract definition. It gives a precise mechanical account of how the dispositions are realised in design terms. It also accords well with the common view of moods as 'emotional *states'*. That is, if emotional episodes are by definition relatively short-lived, this is because such episodes are to be identified, in our model, with the relatively short period during which a lower-layer takes control of behaviour. Whenever, the top layer is in control, the system is not 'in the grip of' an emotion. However, the system is always in some particular emotional state or other, in the sense that we can always give an answer to the question: 'which emotional mechanism has the lowest activation threshold at the moment?' The system is always in one mood or another, even when none of the interruption mechanisms are in control of behaviour. For example, if the anger mechanism has the lowest activation threshold, then we say that the system is in an 'irritable mood'.

*Feedback*

By allowing for feedback effects between the activation thresholds of the interruption mechanisms, we could also model the oscillations of mood that are commonly observed in normal people and which are more pronounced in those suffering from various mood disorders. There might even be a role for coupled oscillators here, which would increase the relevance of a dynamical approach to the mind by showing another role for oscillators in the mental economy other than rhythmic locomotion. Perhaps unipolar mood disorders could be modelled as point attractors in the phase space of emotional states, and bipolar disorders as limit cycles. The dynamics of mood disorders are not well understood, and such a model could prompt us to look for relevant data by suggesting possible control parameters.

*Does the interruption model impute too much internal structure to emotional mechanisms?*

In the previous section, I argued that many criticisms of discrete models can be reduced to the claim that the discrete model imputes too much internal structure to the real-world system. In cognitive science, this amounts to the charge that discrete models sometimes present a rather baroque view of the mind which flouts the principle of parsimony.

It is hard to tell whether or not the interruption theory could be accused of attributing too much internal structure to the interruption mechanisms, since I have not offered any detailed account of how these mechanisms are supposed to work. Pending such accounts, we must suspend our judgement on this question. However, when constructing design hypotheses for these mechanisms, we should bear in mind the general principle that complex behaviour can sometimes be generated by relatively simple mechanisms.

These rather bald statements do not go very far towards demonstrating the usefulness of adopting a dynamical approach to emotion. It remains to be seen, therefore, whether or not the dynamical approach can contribute as much to interruption theory as the other non-classical approaches outlined in this thesis.

## Chapter Six:

# Non-classical cognitive science and emotion

> *In former times a Raja sent for all the blind men in his capital and placed an elephant in their midst. One man felt the head of the elephant, another an ear, another a tusk, another the tuft of its tail. Asked to describe the elephant, one said that the elephant was a large pot, others that it was a winnowing fan, a ploughshare, or a besom. Thus each described the elephant as the part which he first touched, and the Raja was consumed with merriment.*
>
> Buddhist parable[1]

In this final chapter I argue that the three non-classical approaches discussed in the previous chapters may be combined to produce a single integrated non-classical approach. In the second section, I illustrate this approach by reference to the interruption theory of emotion.

### *6.1. Integrated non-classical cognitive science*

In chapter two, I described how, in the period between 1950 and 1980, cognitive scientists tended (i) to assume that the mind was a domain-general mechanism, (ii) to identify its boundary with a physical feature of the agent (either the boundary of the brain, or the boundary of the body), and (iii) to work with discrete models. In chapters three to five, I showed how, after 1980, these assumptions came to be challenged by various 'deviant' schools of thought within the cognitive science community. Evolutionary psychologists argued that the mind was composed of many domain-specific mechanisms. Proponents of the so-called 'situated'

[1] The parable of the elephant is attributed to the Buddha by the *Udana*, one of the scriptures of the Theravada or Hinayana school. This version of the story is taken from *Buddhism: An Introduction and Guide*, by Christmas Humphreys (Harmondsworth: Penguin, 1951), p.11.

approach argued that the mind had flexible boundaries. And there were calls to model the mind in continuous terms by those enamoured of dynamical systems theory.

In this section, I try to lay bare the conceptual links between these three non-classical approaches. First, however, I want to examine the links between the assumptions of the classical approach.

### *The disunity of classical cognitive science*

The three assumptions of domain-generality, internalism and discreteness are logically independent. It is perfectly possible to imagine a cognitive scientist of a classical bent dropping any one of them and retaining the others, or relinquishing any two of them and keeping one. All permutations are possible from a strictly logical point of view. Nor do there seem to be any strong theoretical reasons of a non-logical kind for linking the three together. One is forced to the conclusion, then, that the conjunction of these three assumptions in the first decades of cognitive science was a mere historical accident.

The word 'accident', however, should not be taken to imply that there is no good explanation for the fact that the first generation of cognitive scientists subscribed to these three assumptions. It is just that the explanation should appeal to historical reasons rather than theoretical or logical ones. For example, Andrew Wells has suggested one possible historical explanation for the widespread acceptance of assumption (ii): internalism. He claims that Turing originally conceived of the finite state control and the infinite tape memory in his 'universal machine' as equivalent to the agent and the environment, respectively. According to Wells, it was only the increasing tendency of computer scientists to hardwire the control to a finite memory and package the resulting system into a single box that led Turing's original distinction to be increasingly blurred. The result was that cognitive scientists interpreted control and memory as internal components *within* the cognitive agent rather than as two qualitatively different sources of variance – the organism and the environment (Wells, 1998: 275). In other words, it was a contingent fact about the design of computing machinery that led cognitive scientists to adopt a strong commitment to internalism.

This point might be extended more generally to other aspects of classical cognitive science. The basic idea of CTM is so general that it gives no real guidance about how artificial minds should be designed, or by what criteria they should be evaluated. Yet, as a matter of fact, many classical cognitive scientists probably identified the term 'computer' far too closely with the actual machines of their day, which were built for very particular purposes (and certainly not always to provide models of the human mind, although they were later construed as such). The demands of mathematical rigour from those who wanted machines to ease the burden of doing sums meant that these machines had to be consistent, reliable, non-random, accurate, and predictable.[2] These aims are best met by discrete-state machines. Whether or not this explanation is the correct one remains to be seen, but this would clearly make an interesting project in the history of science.

The unity of classical cognitive science is a historical curiosity, then, that has no interesting theoretical explanation. One might be tempted to infer from this that, just as the three classical assumptions are not conceptually linked, so also their contraries are quite independent of one another. This, however, would not be a valid inference. It might be true, but one cannot assume it on purely logical grounds. In fact, I think it is false. I think that there are good grounds for thinking of the three non-classical approaches as forming a single conceptual bundle. There are good grounds, in other words, for combining the evolutionary, situated and dynamical approaches into a single 'non-classical' approach that is far more theoretically coherent than the classical approach ever was.

[2] This point was brought home to me by Geoffrey Miller (personal communication).

*The unity of non-classical cognitive science*

This is not to say that the non-classical approaches are equal partners. On the contrary, I think that the conceptual unity of the integrated non-classical approach I am proposing comes from taking the evolutionary approach to be primary. In other words, if you start with an evolutionary approach, you will probably also want to adopt a situated approach, and perhaps a dynamical approach too. But if you start with a situated approach you are much freer to retain the classical assumptions of domain-generality and discreteness, and if you start with a dynamical approach, you are not thereby given any guidance on the questions of domain-generality and internalism.

Here is how the evolutionary approach leads us to the other non-classical approaches. First, it leads us to think in situated terms, because natural selection will always favour those cognitive agents that can offload as much of their computational burden onto the environment as possible. Natural selection is an economiser, and computation is expensive. George Williams made a similar point in 1966 when he argued that, other things being equal, natural selection will always favour obligate adaptations over facultative ones (Williams, 1966). Obligate adaptations develop willy-nilly, while facultative ones develop one way in one set of circumstances and another way in other circumstances. Obligate adaptations are computationally cheap; the advantages of polymorphism have to be considerable before natural selection will give up the cheap alternative.

This point is reinforced by computer simulations of evolution. When artificial perceptual discrimination systems are allowed to evolve by techniques of artificial life, the end result is often a highly situated system that offloads as much computation onto the world as possible. In other words, such systems exploit very specific details of their local environment, not general features of all possible environments. For example, a system for picking out squares might simply evolve a straight-edge detector if the only shapes in its artificial environment were squares and circles. Only if it also regularly encountered other straight-edged shapes such as triangles would it need to build in extra knowledge.

*Evolution and dynamics*

The evolutionary approach also leads cognitive scientists to pay attention to the grain of our scales (are there enough states in the model?), to the details of timing, and to chaotic behaviour – all the things which, I argued, are the distinctive traits of the dynamical approach. The details of timing are particularly important to evolved cognitive agents who, unlike the systems that tend to result from traditional human design methods, must deal with a multiplicity of concerns in real time. Evolutionary psychologists are also interested, of course, in longer time-scales than those that apply to the second-by-second control of behaviour. Ontogeny and phylogeny are also inherently temporal processes, so there may be a distinctive advantage in adopting the dynamical approach when trying to understand them.

Ontogeny and phylogeny have traditionally been described in terms that are, at least implicitly, digital. During the past ten years, however, dynamical accounts of these processes have grown in popularity (Depew and Weber, 1995). There are now interesting, though still rather speculative, accounts of both ontogenetic and phylogenetic processes that attribute to them complex nonlinear phenomena such as bifurcation.

Many traits, for example, are classed as innate or acquired, or analysed into innate and acquired components. This binary opposition, however, is too crude for most explanatory purposes. Here, a dynamical approach can help to make it clear that innateness is a question of degree. The resources of dynamical systems theory can also be applied directly to the question of how innate a given trait is by providing a mathematically precise model of Waddington's epigenetic landscape (Waddington, 1940). The various factors influencing the shape of the landscape can each be represented as a different dimension in an *n*-dimensional state space, and the trajectories through this state space will then represent the possible paths taken by individual development or the evolution of a species. Nor is there any need, in such a model, to partition these factors on the basis of whether they are 'genetic' or 'environmental'. Such a distinction serves no useful purpose (other than assisting animal and plant breeders), and the dynamical model makes this clear by treating all factors as equivalent sources of variance.

If one adopts the evolutionary approach to begin with, then, there are good reasons why one should also take a situated and a dynamical approach. The evolutionary approach, therefore, can be the basis of an integrated non-classical approach that differs from the classical approach in taking a view of minds as:

(i) Massively domain-specific
(ii) Leaky
(iii) Inherently temporal, and involving continuous as well as discrete systems

The integrated non-classical approach also differs from the classical approach in that these three tenets are much more closely linked by theoretical reasons than the assumptions of domain-generality, internalism and discreteness. The integrated non-classical approach is, in other words, truly *integrated* in a way that the classical approach is not.

### *From possibility to actuality*

Another difference between the integrated non-classical approach and the classical approach is that, at moment of writing, the former only exists as a theoretical possibility. For several decades, almost all cognitive scientists subscribed to all three assumptions that define the classical approach. Many still do, though the number is gradually diminishing. However, there are very few cognitive scientists who reject all three classical assumptions.

Many cognitive scientists today reject one, and some reject two, but hardly any challenge the classical approach on all three fronts. There are few, if any, proponents of the integrated non-classical approach I propose.

If anything, the proponents of one classical approach often seem to be more concerned to compete with proponents of the others rather than with building links between them. Exchanges between proponents of the various non-classical approaches to cognition can sometimes take on a very shrill tone, as if these approaches were mutually exclusive.

This is less true of relations between the situated and the dynamical approaches, where there have been some attempts to tie them together. Andy Clark, for example, has argued that the resources of dynamical systems theory are strongly preferable for understanding embodied, embedded agents. In particular, he claims that dynamical systems theory provides 'an explanatory framework that (1) is well suited to modelling both organismic and environmental parameters and (2) models them both in a uniform vocabulary and framework, thus facilitating an understanding of the complex interactions between the two' (Clark, 1997: 113). A framework that invokes digital components lacks these advantages.

However, this is far from being the common view among cognitive scientists working with a situated approach or those working under the banner of dynamical systems theory. Most of the time, each of these two schools of thought operate without much reference to the other. The same is true of evolutionary psychology with respect to both the situated and dynamical approaches. One looks in vain at the bibliographies of the various books and papers published by leading evolutionary psychologists for any mention of any works by the leading proponents of situated cognition (such as Andy Clark, Francisco Varela, George Lakoff and Rodney Brooks), nor of the earlier sources on which these writers draw (such as Heidegger, Merleau-Ponty and J. J. Gibson).[3] Likewise, most evolutionary psychologists seem to be unaware of the burgeoning work in dynamical cognition. It 'seems that evolutionary psychologists have, at least up to now, adopted a very narrow conception of cognitive science, one which is exclusively classical in all respects except its commitment to the evolved nature of mind. When reading Cosmides and Tooby, or Steven Pinker, one gets the impression that cognitive science is still entirely dominated by the disembodied, digital approach that characterised the discipline in its early days.

This is particularly unfortunate for, as I have just argued, the evolutionary approach can provide the foundation for an integrated non-classical approach. Evolutionary cognitive scientists are not logically obliged to adopt a situated and a dynamical approach, but there are good theoretical reasons why they should do so. By failing to realise this, evolutionary psychologists have so far missed the opportunity to lead a revolution in cognitive science.

The explanation for this may for have more to do with social and rhetorical reasons than with any failure of imagination on the part of evolutionary psychologists. The main priority for evolutionary psychology in its early days, in the late 1980s, was to establish itself as a credible research program in its own right. At that time, evolutionary theory was perceived as rather tangential to the business of discovering mental structure by most cognitive scientists. In challenging this assumption, as well as arguing for the domain-specificity of many mental processes, evolutionary psychologists had enough on their hands. Tying the evolutionary approach too closely to the situated and dynamical approaches, which were also perceived as 'young Turks' by the older generation of cognitive scientists, would have made it even harder to establish credibility (and thus to get jobs and funding). At a time when departments of cognitive

[3] Geoffrey Miller is somewhat of an exception. He is the only evolutionary psychologist to make abundant references to J. J. Gibson and ecological psychology.

science were run by those of a classical bent, retaining the classical assumptions of internalism and discreteness was probably necessary for evolutionary psychology for purely tactical reasons.[4]

Things are somewhat different now, though. The evolutionary approach is certainly not accepted by all – indeed, there are still sections of the cognitive science community that are vehemently opposed to it – but it has at least established itself as a major player. It has already acquired many of the status symbols of a thriving research program: scholarly journals that are peer-reviewed, learned societies, textbooks, academic positions, undergraduate courses, and annual conferences. Evolutionary psychology has also achieved something that other schools of thought in cognitive science never have, not even the classical approach: widespread popular appeal. This last feature, however, may actually have *hindered* the process of gaining academic acceptance (perhaps because classical cognitive scientists are jealous).

Now that evolutionary cognitive science is no longer an embryonic discipline but a rapidly maturing research program, the time would seem ripe for it to question its adherence to the classical assumptions of internalism and discreteness. If evolutionary psychologists have enough nerve and imagination, they could transform the integrated non-classical approach I have proposed here from a mere theoretical possibility into a historical reality.

### *Pluralism and compatibility*

How would such an integrated non-classical cognitive science relate to its classical forbear? Confrontation would not necessarily be the order of the day. As I have argued in the previous chapters, the cognitive scientist is not necessarily forced to make a stark choice when faced with the issues of domain-generality and discreteness (internalism is a different matter, as

[4] This point was suggested to me by Geoffrey Miller (personal communication).

I will argue shortly). It is quite possible to build hybrid models that combine domain-general mechanisms with domain-specific ones (as Fodor did in *The Modularity of Mind*), or that combine discrete-state machines with analogue systems (e.g. Tani, 1999: 152). Thus, instead of talking about a 'paradigm shift' or a 'scientific revolution', then, the relationship between the integrated non-classical approach and the classical one might therefore be described more appropriately as one of assimilation, akin to the way that the theory of general relativity is sometimes described as absorbing Newton's theory of gravity as a special case applying only to a limited domain. This way of talking might help to avoid perpetuating the spurious confrontations that have dogged cognitive science in recent years. Andy Clark has argued persuasively for an eclectic approach to the mind in which we need to combine a variety of explanatory styles, including both the componential explanations typical of the classical approach and the dynamical explanations of more recent years. He suggests that progress in cognitive science will consist of 'adding new tools' to the explanatory tool-kit, rather than abandoning those we already have. After all, if *the mind* were so simple that a single approach could unlock all its secrets, *we* would be so simple that we couldn't understand the theory![5]

Still, if calls for pluralism are not to descend into sloppy thinking, we must be precise about the nature of the compatibility between the classical approach and the various non-classical approaches. Compatibility comes in various kinds.[6] In particular, the claim that two approaches, A and B, are 'compatible' could be construed in at least three different ways:

(i) A and B provide different kinds of explanation
(ii) A and B are mutually inter-translatable
(iii) A and B explain different phenomena

[5] Adapted from a phrase quoted by Andy Clark (Clark, 1997: 175). Clark states that he was unable to trace the originator of this remark.

[6] I owe this point to Michael Wheeler (personal communication).

I have argued that the evolutionary approach is compatible with the classical approach in sense (i), that the situated approach is compatible with the classical approach in sense (ii), and that the dynamical approach is compatible with the classical approach in sense (iii). I will now spell these compatibility claims out in more detail.

The classical approach and the evolutionary approach provide different *kinds* of explanation. Specifically, the former is fundamentally concerned with providing *design* explanations (how minds work), while the latter is concerned with *functional-historical* explanations (why minds work the way they do). Now, there may be ways of deriving constraints on the former kind of explanation from the latter; this is precisely what Cosmides and Toobey suggest when they propose that cognitive models should always be evaluated according to the evolvability criterion (see section 3.1). However, this possibility has not yet been conclusively demonstrated, as I showed by reference to the debate about domain-specificity. However, even if this were eventually to be proven to be the case, it would not call into question the general point that providing synchronic hypotheses about mental structure and providing diachronic hypotheses about the origins of such structures are different kinds of explanatory project. The links between them, if any, are empirical, not conceptual; we can only derive constraints about mental design from functional explanations by enriching the latter with a whole set of empirical assumptions about the adaptive problems posed by particular environments for particular lineages of organism.

The classical approach and the situated approach are mutually inter-translatable. As I argued in chapter four, whenever the proponent of the situated approach locates the input-output boundary of a particular computational system at some point outside the organism, the classical cognitive scientist can always re-describe this system in traditional internalist terms. Proponents of the situated approach can argue against this internalist move on the grounds that it is dogmatic and is of dubious explanatory value, but they cannot rule it out on purely logical grounds. The leaky mind hypothesis, when stated baldly, is the contradictory of the internalist view, but when it comes to applying these two approaches to real-world systems, there are no purely logical grounds for preferring one over the other. This logical compatibility, however, is not that exciting.

Finally, the classical approach and the dynamical approach are compatible in sense (iii): they explain different phenomena (and sometimes different aspects of the same phenomenon, which amounts to the same thing). I argued in chapter five that, insofar as continuous models can be taken as a separate class, they are better suited than discrete models to explaining 'low-level' processes such as pattern-matching and limb-coordination. Discrete models, on the other hand, are better suited to modelling 'high-level' processes such as forward planning. Hence the cognitive scientist is not forced to choose between an exclusively discrete approach and an exclusively continuous one. She can construct hybrid models that use both digital and analogue components. The same point applies to the debate about domain-generality. It is possible to conceive of minds that employ both domain-general mechanisms and domain-specific ones. This, indeed, was the main thrust of Fodor's proposal in *The Modularity of Mind* (Fodor, 1983).

Calls for a kind of *super-integrated* cognitive science, combining the insights of the integrated non-classical approach and the classical one, must be careful to distinguish between these different kinds of compatibility. Otherwise, they risk leading us to blur the questions raised by the non-classical approaches, and thus to obscure their importance. It is hard to argue with calls for broad-mindedness and pluralism, but if such calls are to amount to anything more than the politically-correct view that 'everyone must have prizes', they must tempt us into thinking that there are no real disagreements. Methodological pluralism is not an end in itself; it is simply a way of clearing aside the false oppositions so that we may concentrate on the genuine ones. The ultimate aim of cognitive science, after all, should not be to provide a cosy umbrella under which those of any persuasion can take shelter, but to answer the questions about how minds work and why they work the way they do.

## *6.2. An integrated non-classical approach to emotion*

In section 3.2, I put forward a theory of emotion based on a proposal by Herbert Simon. According to this theory, emotions are defined in functional terms as interruption mechanisms. An interruption mechanism is one that can interrupt ongoing activity and temporarily take control of behaviour in the service of survival or reproductive goals. In section 4.2, I developed the interruption theory by showing how some of the computational burden it required could be offloaded from the brain onto the body and the environment. In section 5.2, I asked whether further refinements to the theory could be made by paying attention to the continuous features of emotion as well as the discrete ones.

The interruption theory can serve as an example of the integrated non-classical approach that I proposed in the previous section. It has all the hallmarks that define such an approach: domain-specificity, leakage, and continuity. Furthermore, the evolutionary approach is primary. The view of emotions as interruption mechanisms is based on the functional accounts of emotion provided by evolutionary psychology. Thus I started with evolutionary considerations, and used the situated and dynamical approaches serve to round out the theory.

The interruption theory, as I have sketched it out in this thesis, needs much more conceptual refinement. However, the bare bones are at least clear. The theory is already capable of generating specific hypotheses about the design of emotional systems. Such hypotheses could be tested by implementing these designs in artificial agents. Indeed, without this vital step, the interruption theory would remain of limited use to cognitive science.

### *Classical versus non-classical approaches to emotion*

The interruption theory is clearly non-classical in flavour, but this does not mean that it is incompatible with classical models of emotion. In the previous section, I outlined various ways in which theories can be compatible: providing different kinds of explanation, being mutually-intertranslatable, and explaining different things. Is the interruption theory compatible with classical appraisal theory in any (or all) of these different ways?

#### *(i) Kinds of explanation*

The interruption theory provides a very different *kind* of explanation to that provided by the classical account of emotion. Appraisal theory and the propositional attitude theory both explain *how* emotions can be representational. The interruption theory takes this as given, and draws on evolutionary considerations to explain what emotions are representations *of*. Emotions, it claims, are representations of changes that are relevant to the achievement of some biological (i.e. survival-related or reproduction-related) goal. To be more precise, it is the function of emotions to represent such changes to the organism so that behaviour can be modified accordingly.

This general account needs to be fleshed out for each individual emotion. For each emotion, we need to say which biological goal the emotion is designed to serve. This provides a criterion for individuating emotional mechanisms: one goal, one mechanism. For example, fear is the emotion designed to serve the goal of avoiding potential physical injury (trauma), whether as a result of being attacked by another organism, or as a result of some process in the nonliving environment such as an avalanche. The function of fear is thus to represent this danger to the organism so that ongoing activity can be interrupted, when necessary, to avoid the danger.

The interruption theory complements the classical approach because the functional account it provides offers a new way of distinguishing between cognition and emotion. The classical approach had a rather conservative aim; to show how emotions could be seen as thoroughly representational, and thus count as true mental phenomena according to CTM. Appraisal theory and the propositional attitude theory achieved this aim by construing emotions as judgements. But this success was purchased at the price of destroying the only widely accepted account of how emotions differed from thoughts. Prior to appraisal theory and the propositional attitude theory, most psychologists had accepted Hume's view that passion was to be distinguished from reason by reference to the concept of representation; reason was representational, while passion was not. The classical approach rejects this way of making the distinction between cognition and emotion, but offers no other way to make it. The interruption theory does. If emotions are interruption mechanisms, then, if cognition is a distinct kind of process, it must be the kind of process that cannot interrupt any other.

### *(ii) Inter-translatability*

When the interruption theory is supplemented with considerations drawn from the situated approach, emotions can be seen as processes occurring in a system whose boundaries are not co-terminous with any single feature of the organism. On this view, although it is sometimes useful to view emotions as entirely neural processes, at other times it is useful to see them as processes that leak out of the brain into the rest of the body and even into the external world. Humans, in particular, have found ways of offloading the computational burden involved in deciding when to interrupt ongoing activity onto parts of their environment. Other animals succeed in doing this too, though to a lesser extent.

From a strictly logical point of view, there is nothing to prevent the classical cognitive scientist from re-describing the examples of embodied and embedded emotions in terms of a purely neural process that has complex feedback loops with bodily and environmental processes. The input-output boundary can always be moved back to the borders of the brain. Yet this type of compatibility claim is purely scholastic. There is more of an argument to be had here when one compares the two approaches with regard to explanatory fruitfulness rather than mere logical consistency. The design-based methodology of cognitive science means that the proper way to assess a research program is not via post-hoc logical assimilation to a given theory, but via practical research. In other words, the mere fact that two approaches are mutually intertranslatable should not lead us to overlook the possibility that one approach might lead consistently to much better working models.

### *(iii) Explanatory domains*

If dynamical considerations can enrich the interruption theory by leading to a greater attention to the temporal features of emotion, this does not necessarily lead to any incompatibility with the classical approach. The classical models of emotion may not have been sufficiently sensitive to the details of timing, but nor did they deny the importance of such details. They simply ignored them. In this case, the interruption theory can be seen as supplementing the classical approach to emotion by providing explanations of different kinds of emotional phenomena. There is no real conflict here.

### *Domain-generality*

The fact that the interruption theory is compatible with the classical approach to emotion in all these ways should not tempt us into thinking that there are no areas of disagreement at all. There does seem to be at least one major point of disagreement – the issue of domain-generality. In the classical models, such as the OCC model, all emotional stimuli are processed by a single emotion-generating mechanism. In the interruption theory, however, each emotion is implemented by a distinct mechanism that attends only to the kind of input that is relevant to that emotion.

However, it would be false to infer from this that the interruption theory is incompatible with the classical approach on this point. Although the actual models of emotion developed by appraisal theorists have all been domain-general, this need not have been the case. The domain-generality of appraisal-type models is a historical accident, not a theoretical necessity. There is nothing *inherently* domain-general about appraisal theory. The basic idea that emotions are judgements arrived at by attending to particular features of external (and possibly internal) stimuli is compatible with both a domain-general and a domain-specific theory of emotional mechanisms. Thus it is not quite correct to say that the interruption theory is opposed to the classical approach on this point. It is more accurate to say that the interruption theory is opposed to most of the models designed by appraisal theorists so far. Besides, taking a domain-specific view of emotion does not imply that one is also committed to a domain-specific view of cognition. It is possible to conceive of a hybrid model of the mind in which emotional processes are domain-specific while cognitive processes are domain-general.

### *An integrated theory of emotion*

The interruption theory can be seen then, not just as an integrated *non-classical* approach to emotion, but as an integrated approach *simpliciter*. To call it 'non-classical' implies a wholesale rejection of the ideas of domain-generality, internalism and discreteness, taken as a single package. The interruption theory, however, is much more sophisticated than this. As I have just argued, it is compatible (in various different ways) with many aspects of the classical approach to emotion, and many aspects of the classical approach to cognition. To describe the interruption theory as 'non-classical' would be to obscure this compatibility.

This is all well and good. But since understanding emotions, at least from a cognitive perspective, entails building emotional machines, the litmus test of the interruption theory will not rest on such compatibility claims but on whether or not it leads us to better machine models of emotion. If it does, then this would not only vindicate the interruption theory itself, but would also support my proposal for an integrated approach to cognitive science in general.

# *Bibliography*

Abelson, R. P. (1963). 'Computer simulation of 'hot' cognition' in Computer Simulations of Personality. S. S. Tomkins and S. Messick. New York, Wiley.

Ariew, A. (1996). “Innateness and canalization.” Philosophy of Science **63**(Proceedings): S19-S27.

Arnold, M. B. (1960). Emotion and Personality. New York, Columbia University Press.

Arnold, M. B. (1968). 'Introduction' in The Nature of Emotion. M. B. Arnold. Harmondsworth, Penguin: 9-14.

Auyang, S. (1998). Foundations of Complex Systems Theories - in Economics, Evolutionary Biology, and Statistical Physics. Cambridge, Cambridge University Press.

Axelrod, R. (1984). The Evolution of Co-operation. Harmondsworth, Penguin.

Bak, P. (1996). How Nature Works: The Science of Self-Organized Criticality. Berlin, Springer.

Bateson, P. (1991). 'Are there principles of behavioural development?' in The Development and Integration of Behaviour. P. Bateson. Cambridge, Cambridge University Press.

Bateson, P. and P. Martin (1999). Design for a Life: How Behaviour Develops. London, Jonathan Cape.

Beck, A. T., A. J. Rush, et al. (1979). Cognitive Therapy of Depression. New York, Guilford.

Brentano, F. (1874). Psychology from an Empirical Standpoint. London, Routledge and Kegan Paul, 1973.

Brooks, R. A. (1991). 'Intelligence without representation' in Mind Design II: Philosophy, Psychology, Artificial Intelligence. J. Haugeland. Cambridge, Mass. and London, MIT Press: 395-420.

Cahn, J. E. (1990). “The generation of affect in sythesized speech.” Journal of the American Voice I/O Society **8**: 1-19.

Canamero, D. (1997). 'Modeling motivations and emotions as a basis for intelligent behavior' in Proceedings of the First International Conference on Autonomous Agents. W. L. Johnson. New York, ACM Press: 148-155.

Cannon, W. B. (1927). 'The James-Lange theory of emotion' in The Nature of Emotion. M. B. Arnold. Harmondsworth, Penguin, 1968: 43-52.

Caporael, L. R. (1989). “Mechanisms matter: The difference between sociobiology and evolutionary psychology.” Behavioral and Brain Sciences **12**: 17-18.

Chomsky, N. (1959). “Review of B. F. Skinner's *Verbal Behavior.*” Language **35**: 26-58.

Chomsky, N. (1986). Knowledge of Language. New York, Praeger.

Chomsky, N. (1988). Language and Problems of Knowledge: The Managua Lectures. Cambridge, Mass., MIT Press.

Clark, A. (1997). Being There: Putting Brain, Body, and World Together Again. Cambridge, MA and London, MIT Press.

Cole, J. A. (1998). About Face. Cambridge, MA: MIT Press.

Cosmides, L. and J. Tooby (1987). 'From evolution to behavior: evolutionary psychology as the missing link' in The Latest on the Best: Essays on Evolution and Optimality. J. Dupre. Cambridge, MA, MIT Press: 277-306.

Cowie, F. (1998). "Mad dog nativism." The British Journal for the Philosophy of Science **49**(2): 227-252.

Cummins, D. D. and R. Cummins (1999). "Biological preparedness and evolutionary explanation (authors' manuscript)." .

Damasio, A. R. (1994). Descartes' Error: Emotion, Reason and the Human Brain. London, Papermac.

Darwin, C. (1872). The Expression of the Emotions in Man and Animals. Chicago, University of Chicago Press, 1965.

Dawkins, R. (1982). The Extended Phenotype: The Long Reach of the Gene. Oxford, Oxford University Press.

Dawkins, R. (1986). The Blind Watchmaker. Harmondsworth, Penguin.

de Sousa, R. (1987). The Rationality of Emotion. Cambridge, MA, MIT Press.

Dennett, D. C. (1990). 'Cognitive wheels: the frame problem of AI' in The Philosophy of Artificial Intelligence. M. Boden. Oxford, Oxford University Press.

Dennett, D. C. (1995). Darwin's Dangerous Idea: Evolution and the Meanings of Life. Harmondsworth, Penguin.

Dennett, D. C. (1996). Kinds of Minds: Toward an Understanding of Consciousness. New York, Basic Books.

Dennett, D. C. (1999). The zombic hunch: extinction of an intuition? Royal Institute of Philosophy Millenial Lecture, London.

Depew, D. J. and B. H. Weber (1995). Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection. Cambridge, Mass. and London, England, MIT Press.

Dreyfus, H. L. (1979). 'From micro-worlds to knowledge representation: AI at an impasse' in Mind Design II: Philosophy, Psychology, Artificial Intelligence. J. Haugeland. Cambridge, Mass. and London, MIT Press, 1997: 143-182.

Driver, P. and D. Humphries (1988). Protean Behaviour: The Biology of Unpredictability. Oxford, Oxford University Press.

Edelman, G. (1992). Bright Air, Brilliant Fire: On the Matter of the Mind. Harmondsworth, Penguin.

Ekman, P. (1992). “An argument for basic emotions.” Cognition and Emotion **6**(3/4): 169-200.

Ekman, P., R. W. Levenson, et al. (1983). “Autonomic nervous system activity distinguishes among emotions.” Science **221**(4616): 1208-10.

Eldredge, N. and S. J. Gould (1972). 'Punctuated equilibria: an alternative to phyletic gradualism' in Models in Paleobiology. T. J. M. Schopf. San Francisco, Freeman, Cooper and Company: 82-115.

Elliot, C. (1994). "Components of two-way emotion communication between humans and computers using a broad, rudimentary model of affect and personality." Cognitive Studies: Bulletin of the Japanese Cognitive Science Society **1**(2): 16-30.

Elman, J. L. and E. A. Bates (1996). Rethinking Innateness: A Connectionist Perspective on Development. Cambridge, MA, MIT Press.

Essa, I. and A. Pentland (1997). "Coding, analysis, interpretation and recognition of facial expressions." IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7): 757-763.

Etcoff, N. L. and J. J. Magee (1992). "Categorical perception of facial expressions." Cognition **44**: 227-240.

Fodor, J. A. (1968). 'The appeal to tacit knowledge in psychological explanation' in Mind and Cognition: An Anthology, Second Edition. W. G. Lycan. Oxford, Blackwell: 46-49.

Fodor, J. A. (1975). The Language of Thought. Cambridge, Mass., Harvard University Press.

Fodor, J. A. (1980). "Methodological solipsism considered as a research strategy in cognitive psychology." Behavioral and Brain Sciences **3**: 63-73.

Fodor, J. A. (1984). "Observation reconsidered." Philosophy of Science **51**: 23-43.

Fodor, J. A. (1987). 'Modules, frames, fridgeons, sleeping dogs, and the music of the spheres' in Modularity in Knowledge Representation and Natural-Language Understanding. J. Garfield. Cambridge, MA, MIT Press: 26-36.

Fodor, J. A. (2000). The Mind doesn't Work that Way. Cambridge, MA, MIT Press.

Fodor, J. A. and Z. W. Pylyshyn (1988) "Connectionism and cognitive architecture: a critical analysis", Cognition **28**: 3-71.

Frank, R. H. (1988). Passions within Reason: the Strategic Role of the Emotions. New York, Norton.

Frijda, N. (1986). The Emotions. Cambridge, Cambridge University Press.

Gardner, H. (1987). The Mind's New Science: A History of the Cognitive Revolution. New York, Basic Books.

Goleman, D. (1995). Emotional Intelligence. New York, Bantam Books.

Gould, S. J. (1983). 'The hardening of the Modern Synthesis' in Dimensions of Darwinism. M. Greene. New York, Cambridge University Press.

Graham-Rowe, D. (1999) Emotional machinations. Daily Telegraph: 10.

Gray, R. D. (1992). 'Death of the gene: developmental systems strike back' in Trees of Life: Essays in the Philosophy of Biology. P. E. Griffiths. Dordrecht, Kluwer: 165-209.

Griffiths, P. E. (1990). 'Modularity and the psychoevolutionary theory of emotion' in Mind and Cognition: An Anthology. W. G. Lycan. Oxford, Blackwell, 1999.: 516-529.

Griffiths, P. E. (1997). What Emotions Really Are: The Problem of Psychological Categories. Chicago & London, University of Chicago Press.

Griffiths, P. E. (1999). "Author's response." Metascience **8**(1): 49-59.

Harnad, S. (1999) Turing on reverse-engineering the mind.

Haugeland, J. (1996). 'What is mind design?' in Mind Design II: Philosophy, Psychology, Artificial Intelligence. J. Haugeland. Cambridge, Mass. and London, England, MIT Press, 1997.

Heidegger, M. (1926). Being and Time: A Translation of *Sein und Zeit*. Albany, SUNY Press, 1996.

Hofstadter, D. R. and D. C. Dennett, Eds. (1981). The Mind's I: Fantasies and Reflections on Self and Soul/ composed and arranged by Douglas R. Hofstadter and Daniel C. Dennett. Brighton, Harvester Press.

Hume, D. (1734). A Treatise on Human Nature. Oxford, Clarendon Press, 1955.

Humphrey, N. (1992). A History of the Mind. New York, Simon and Schuster.

Humphrey, N. (1997). "A review of *Kinds of Minds* by Daniel C. Dennett." The Journal of Philosophy **94**: 97-103.

Hutchins, E. (1995). "How a cockpit remembers its speeds." Cognitive Science **19**(3): 265-288.

Izard, C. E. (1979) "Emotions as motivations: an evolutionary developmental perspective" in Nebraska Symposium on Motivation Vol. 27 ed. R. Dienstbier. Lincoln: University of Nebraska Press.

Izard, C. E. (1993). "Four systems for emotion activation: cognitive and noncognitive processes." Psychological Review **100**(1): 68-90.

James, W. (1884). 'What is an emotion?' in The Nature of Emotion. M. B. Arnold. Harmondsworth, Penguin, 1968: 17-36.

Karmiloff-Smith, A. (1992). Beyond Modularity: A Developmental Perspective on Cognitive Science. Cambridge, MA, MIT Press.

Karmiloff-Smith, A. (1999). 'Modularity of mind' in The MIT Encyclopedia of the Cognitive Sciences. R. A. Wilson and F. C. Keil. Cambridge, MA and London, England, MIT Press: 558-560.

Kauffman, S. (1995). At Home in the Universe: the Search for Laws of Complexity. Harmondsworth, Penguin.

Kelso, J. A. S. (1995). Dynamic Patterns: The Self-Organisation of Brain and Behaviour. Cambridge, MA, and London, MIT Press.

Koda, T. (1996) Agents with faces: a study on the effects of personification of software agents. MIT Media Lab. Cambridge, Massachussetts Institute of Technology.

Krebs, J. R. and R. Dawkins (1984). 'Animal signals, mind-reading and manipulation' in Behavioural Ecology: An Evolutionary Approach, Second Edition. J. R. Krebs and N. B. Davies. Oxford, Blackwell Scientific: 380-402.

Kuhn, T. (1962). The Structure of Scientific Revolutions. Chicago, University of Chicago Press, Third Edition, 1996.

Lakatos, I. (1970). 'Falsification and the methodology of scientific research programmes' in Criticism and the Growth of Knowledge. I. Lakatos and A. Musgrave. Cambridge, Cambridge University Press.

Lakoff, G. and M. Johnson (1999). Philosophy in the Flesh: the Embodied Mind and its Challenge to Western Thought. New York, Basic Books.

Langton, C. G. (1986). "Self-reproduction in cellular automata." Physica D **10**: 1120-1149.

Lazarus, R. S. (1991). Emotion and Adaptation. New York, Oxford University Press.

LeDoux, J. (1998). The Emotional Brain: The Mysterious Underpinnings of Emotional Life. London, Weidenfeld & Nicholson.

Levenson, R. W., P. Ekman, et al. (1990). "Voluntary facial action generates emotion-specific autonomic nervous system activity." Psychophysiology **27**(4): 363-384.

Lormand, E. (1999). 'Frame problem' in The MIT Encyclopedia of the Cognitive Sciences. R. A. Wilson and F. C. Keil. Cambridge, Mass. and London, MIT Press: 326-327.

MacLean, P. D. (1973). A Triune Concept of the Brain and Behaviour. Toronto, University of Toronto Press.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco, Freeman.

Maynard Smith, J. (1982). Evolution and the Theory of Games. Cambridge, Cambridge University Press.

McCulloch, W. S. (1965). Embodiments of Mind. Cambridge, MA, MIT Press.

McCulloch, W. S. and W. Pitts (1943). "A logical calculus of the ideas immanent in nervous activity." Bulletin of Mathematical Biophysics **5**(115-133).

Miller, G. and D. Cliff (1994). 'Protean behaviour in dynamic games: arguments for the co-evolution of pursuit-evasion tactics in simulated robots' in From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behaviour. D. Cliff, P. Husbands, J. A. Meyer and S. Wilson. Cambridge, MA, MIT Press: 411-420.

Miller, G. F. (1997). 'Protean primates: the evolution of adaptive unpredictability in competition and courtship' in Machiavellian Intelligence II: Extensions and Evaluations. A. Whiten and R. Byrne. Cambridge, Cambridge University Press: 312-340.

Moffat, D., N. Frijda, et al. (1994) Analysis of a model of emotions, Cogprints.

Murphy, D. and S. Stich (1998). "Darwin in the madhouse: Evolutionary psychology and the classification of mental disorders (unpublished manuscript)." .

Neisser, U. (1967). Cognitive Psychology. New York, Appleton-Century-Crofts.

Nesse, R. (1990). "Evolutionary explanations of emotions." Human Nature **1**(3): 261-289.

Newell, A. and H. A. Simon (1976). 'Computer science as empirical enquiry: symbols and search' in Mind Design: Philosophy, Psychology, Artificial Intelligence (1997). J. Haugeland. Cambridge, Mass. and London, England, MIT Press: 81-110.

Nowlis, V. (1963) 'The concept of mood' in Conflict and Creativity, ed.S. M. Farber and R. H. L. Wilson. New York: McGraw-Hill.

Oatley, K. (1999). 'Emotions' in The MIT Encyclopedia of the Cognitive Sciences. R. A. Wilson and F. C. Keil. Cambridge, Mass. and London, MIT Press: 273-275.

Oatley, K. and J. M. Jenkins (1996). Understanding Emotions. Oxford and Cambridge, MA, Blackwell.

Ortony, A., G. L. Clore, et al. (1988). The Cognitive Structure of Emotions. Cambridge, Cambridge University Press.

Oyama, S. (1985). The Ontogeny of Information. Cambridge, Cambridge University Press.

Picard, R. W. (1997). Affective Computing. Cambridge, Mass. and London, England, MIT Press.

Pinker, S. (1994). The Language Instinct. Harmondsworth, Penguin.

Pinker, S. (1997). How the Mind Works. Harmondsworth, Penguin.

Plutchik, R. (1980). Emotion: A Psychoevolutionary Synthesis. New York, Harper and Row.

Pugmire, D. (1998). Rediscovering Emotion. Edinburgh, Edinburgh University Press.

Putnam, H. (1960). 'Minds and machines' in Dimensions of Mind: A Symposium. S. Hook. New York, New York University Press.

Pylyshyn, Z., Ed. (1987). The Robot's Dilemma. Norwood, NJ, Ablex.

Quartz, S. R. and T. J. Sejnowski (1997). “The neural basis of cognitive development: a constructivist manifesto.” Behavioral and Brain Sciences **20**(4): 537-596.

Ray, T. S. (1992). 'An approach to the synthesis of life' in The Philosophy of Artificial Life. M. A. Boden. Oxford, Oxford University Press: 111-145.

Roberts, R. (1988). "What an emotion is: a sketch." The Philosophical Review **79**: 183-209.

Roseman, I. J., A. A. Antoniou, et al. (1996). "Appraisal determinants of emotions: constructing a more accurate and comprehensive theory." Cognition and Emotion **10**(3): 241-277.

Roy, D. and A. Pentland (1996). Automatic spoken affect analysis and classification. The Second International Conference on Automatic Face and Gesture Recognition, Killington, VT.

Salovey, P. and J. D. Mayer (1990). "Emotional intelligence." Imagination, Cognition and Personality **9**(3): 185-211.

Samuels, R. (1998). "Evolutionary psychology and the massive modularity hypothesis." The British Journal for the Philosophy Science **49**: 575-602.

Samuels, R. (1998). "What brains won't tell us about the mind: a critique of the neurobiological argument against representational nativism." Mind and Language **13**(4): 548-570.

Scales, J. A. and R. Snieder (1999). "What is a wave?" Nature **401**(6755): 739-740.

Schachter, S. (1964). 'The interaction of cognitive and physiological determinants of emotional state' in Advances in Experimental Social Psychology. L. Berkowitz. New York, Academic Press. **1:** 49-80.

Schachter, S. and J. E. Singer (1962). "Cognitive, social, and physiological determinants of emotional state." Psychological Review **69**: 379-99.

Scholl, B. J. and A. M. Leslie (1999). “Modularity, development and 'theory of mind'.” Mind and Language **14**(1): 131-153.

Segal, G. (1996). 'The modularity of theory of mind' in Theories of Theories of Mind. P. Carruthers and P. Smith. Cambridge, Cambridge University Press: 141-157.

Seifert, C. M. (1999). 'Situated cognition and learning' in The MIT Encyclopedia of the Cognitive Sciences. R. A. Wilson and F. C. Keil. Cambridge, Mass. and London, MIT Press.

Seligman, M. E. P. (1975). Helplessness: On Depression, Development and Death. San Francisco, W. H. Freeman.

Simon, H. A. (1967). “Motivational and emotional controls of cognition.” Psychological Review **74**: 29-39.

Smith, C. A. and P. C. Ellsworth (1985). “Patterns of cognitive appraisal in emotion.” Journal of Personality and Social Psychology **56**: 339-353.

Smith, P. (1998). Explaining Chaos. Cambridge, Cambridge University Press.

Solomon, R. (1977). The Passions. New York, Doubleday.

Sperber, D. (1994). 'The modularity of thought and the epidemiology of representations' in Mapping the Mind: Domain Specificity in Cognition and Culture. L. A. Hirschfeld and S. A. Gelman. Cambridge, Cambridge University Press: 39-67.

Sterelny, K. and P. E. Griffiths (1999). Sex and Death: An Introduction to Philosophy of Biology. Chicago, University of Chicago Press.

Stewart, I. (1989). Does God Play Dice? The New Mathematics of Chaos. Harmondsworth, Penguin, second edition, 1997.

Stich, S. P. (1983). From Folk Psychology to Cognitive Science. Cambridge, Mass. and London, MIT Press.

Suchman, L. A. (1987). Plans and Situated Action. Cambridge, Cambridge University Press.

Swagerman, J. (1987) The Artificial Concern REalization System ACRES: A computer model of emotion. Amsterdam, University of Amsterdam.

Tani, J. (1999) 'An interpretation of the "self" from the dynamical systems perspective: a constructivist approach', in Models of the Self, eds Shaun Ghallagher and Jonathan Shear, Thorverton: Imprint Academic, pp.149-176.

Toda, M. (1962). "Design of a fungus-eater." Behavioral Science **7**: 164-183.

Tomkins, S. (1970). 'Affect as the primary motivational system' in Feelings and Emotions: The Loyola Symposium. M. B. Arnold. New York, Academic Press: 101-110.

Tomkins, S. S. (1982). 'Affect theory' in Emotion in the Human Face. P. Ekman. Cambridge, Cambridge University Press: 353-95.

Tooby, J. and L. Cosmides (1990). "The past explains the present: emotional adaptations and the structure of ancestral environments." Ethology and Sociobiology **11**: 375-424.

Tooby, J. and L. Cosmides (1992). 'The psychological foundations of culture' in The Adapted Mind: Evolutionary Psychology and the

Generation of Culture. J. L. Barkow, L. Cosmides and J. Tooby. New York, Oxford University Press.

Turing, A. M. (1937). "On computable numbers, with an application to the Entscheidungsproblem." Proceedings of the London Mathematical Society **42**: 230-265.

Turing, A. M. (1950). 'Computing machinery and intelligence' in Mind Design II: Philosophy, Psyhology, Artificial Intelligence. J. Haugeland. Cambridge, Mass. & London, England, MIT Press, 1997**:** 29-56.

Turing, A. M. (1952). "The chemical basis of morphogenesis." Philosophical Transactions of the Royal Society **B237**: 37-52.

Turkle, S. (1984). The Second Self: Computers and the Human Spirit. New York, Simon and Schuster.

van Gelder, T. (1998). "The dynamical hypothesis in cognitive science." Behavioural and Brain Sciences **21**: 615-665.

Varela, F. J., E. Thompson, et al. (1991). The Embodied Mind: Cognitive Science and Human Experience. Cambridge MA and London, MIT Press.

Velasquez, J. (1996) Cathexis: A computational model for the generation of emotions and their influence in the behavior of autonomous agents. Media Lab. Massachussetts, MIT.

Waddington, C. H. (1940). Organisers and Genes. London, Allen and Unwin.

Waddington, C. H. (1957). The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology. London, Ruskin House/George Allen and Unwin Ltd.

Weinberg, S. (1993). Dreams of a Final Theory: The Search for the Fundamental Laws of Nature. London, Vintage.

Wells, A. J. (1998). "Turing's analysis of computation and theories of cognitive architecture." Cognitive Science **22**(3): 269-294.

Wheeler, M. (1996). 'From robots to Rothko: the bringing forth of worlds' in The Philosophy of Artificial Life. M. A. Boden. Oxford, Oxford University Press: 209-236.

Williams, G. C. (1966) Adaptation and Natural Selection. Princeton, NJ: Princeton University Press.

Wilson, R. A. (1999). 'Individualism' in The MIT Encyclopedia of the Cognitive Sciences. R. A. Wilson and F. C. Keil. Cambridge, Mass. & London, England, MIT Press: 397-399.

Wright, I. and A. Sloman (1996) MINDER1: an implementation of a proto-emotional agent architecture. Birmingham, University of Birmingham.

Zajonc, R. B. (1980). "Feeling and thinking: preferences need no inferences." American Psychologist **35**: 151-175.