

## **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.<sup>1</sup> 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,

---

<sup>1</sup> 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.

claiming 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).<sup>2</sup> 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 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

---

<sup>2</sup> 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'.

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.<sup>3</sup>

---

<sup>3</sup> 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.

*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 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 neuronanatomy, 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 possess'd 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 cognitive 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 realizability thesis applies just as much to emotion as to other kinds of mental process. In other words, once 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.