Chapter 2
The Computational Mind

2.1 The Computer Analogy

The computational theory of mind grows out of the conception of the brain as an information-processing device, analogous to a computer. In comparison with earlier analogies—brain as hydraulic mechanism, as steam engine, as telephone switchboard—the computer analogy has been remarkably successful in capturing the general public's imagination (see Turkle 1984) as well as in generating fruitful programs of research.

Two properties of computers recommend the analogy. First, the information content of data and programs (especially those written in high-level languages) can be stated independently of physical instantiation in any particular computer. For example, a FORTRAN program will run on more or less any machine, whether built out of vacuum tubes, transistors, or chips. Thus there is a sense in which, like the mind, the information in the computer is autonomous—inhabits a separate domain—from the (mere) hardware that supports computation. Second, the ways in which programs are organized—in terms of goals and subgoals, self-monitoring and self-modification—resonate with commonsense intuitions about the organization of problem solving, learning, and other cognitive tasks. (In addition, the speed and accuracy of computers at tasks like calculation and graphics generation cannot be followed in real time by humans and hence can lead to an impression of intelligence.)

Very crudely, then, the computer analogy suggests the following hypothesis: just as we need not deal with the actual wiring of the computer when writing our programs, so we can investigate the information processed by the brain and the computational processes the brain performs on this information, independent of questions of neurological implementation. This approach is often called *functionalism*; the idea behind this term is that the function rather than the physical substance of the brain is significant in studying the mind.

Such inquiry has occupied much of cognitive psychology and linguistics since the early 1950s. And although it was common even earlier to speak of computers as "electronic brains," the earliest reference I have come

across that makes explicit the analogy to *minds* is Putnam's "Mind and Machines" (1960). It is now routine to speak of the information in the brain as *mental representations* and of the processes operating on such representations as *mental processes*. In short, the mind is taken to stand to the brain as the software and data of the computer stand to the hardware.

2.2 Attractions of the Computational Mind; *Theory I*

Let us see how well this computational notion of the mind corresponds to the traditional phenomenological notion. To begin with, both kinds of mind seem to inhabit a domain of description separate from that of the physical device, though both are causally dependent on the characteristics and states of the physical device. Moreover, thinking in computational terms is a fruitful way to achieve the sort of abstractness and generality that seems necessary to describe not only our conscious life but also the behavior of animals of any substantial degree of complexity (see section 3.1.) In particular, the units of a high-level computational account, such as objects, organized actions, and goals, correspond much more closely to the intuitive units of consciousness than do any presently known descriptions of neural activity.

The computer analogy also offers an attractive way to treat the notion of the unconscious mind. In chapter 1 we saw that the unconscious offers a paradox to the definition of the mental as that which is experienced: how can something be an experience and yet *un*experienced? However, if "mental" is defined in terms of information processing instead, a rather satisfying result emerges. Consider memory, for example. A computer stores a great deal of information, only some small part of which is being actively used at a given time. Similarly, the moment-by-moment kaleidoscopic shifts in information being actively processed in the brain correspond intuitively to the ever-changing stream of consciousness, whereas inactive stored information can be regarded as unconscious—but still mental.

As Putnam (1960) points out, the computer analogy also helps explain how one can know something without being aware that one knows it. Suppose that in order to know some fact *F*, a machine must contain some configuration of computational states *C*. Then, for the machine to be *aware* of knowing *F*, it must be *aware of being in configuration C*. But that requires the realization of some further computational state *C*' that checks whether *C* is present. If *C* is present without *C*' the machine knows *F* but is not aware of knowing *F*. That is, the computational device's self-monitoring is a set of processes beyond those responsible for ordinary interaction with the world. We thus see the emergence of a distinction like that between primary and reflective awareness.

A computer may carry on several tasks at once, some of which involve higher-level goals or forms of organization and some of which are more subsidiary. Since the same is true of the brain (in fact, much more so—see section 3.2), it is tempting to divide the computational mind up into black boxes and identify some especially important one as the locus of consciousness. Then one's hypothesis, roughly, is that information and/or processing that is active within this particular black box is conscious and the rest is unconscious. Thus, the conscious and the unconscious parts of the mind are of the same essential character, built out of information and the processes that operate on it.

Let us codify this approach as Theory I, the first of a series of theories we will develop in an attempt to gradually refine the problem of consciousness.

> *Theory I*
> The elements of conscious awareness consist of information and processes of the computational mind that (1) are active and (2) have other (as yet unspecified privileged properties.

Something like Theory I is common in the literature, in many disparate guises. Minsky (1968) sees consciousness as the "supreme organizer" that can access and debug other faculties; Dennett (1969) identifies the elements of consciousness as the contents of the speech center; Johnson-Laird (1983, 465) says, "The contents of consciousness are the current values of parameters governing the high-level computations of the operating system"; Frith (1981) speaks of consciousness as a monitor system, a higher-level system that directs subroutines, controlling and selecting; Mandler (1984, 89) wants to think of the contents of consciousness as mental products that are undergoing a particular mode of processing. All of these are in a sense anticipated by William James (1890, 288):

> Consciousness consists in the comparison of [simultaneous possibilities] with each other, the selection of some, and the suppression of the rest by the reinforcing and inhibiting agency of attention. The highest and most elaborated mental products are filtered from the data chosen by the faculty next beneath, out of the mass offered by the faculty below that, which mass in turn was sifted from a still larger amount of simpler material, and so on.

The disagreements among these views lie largely in how to spell out proviso (2) in Theory I—what part of the active mind is privileged. One thing they have in common, though, is their unquestioned identification of consciousness with an especially high-level representation or process. This choice corresponds to the traditional intuition that the conscious mind is connected with the will, with the initiation and coordination of action, with the ability to make rational choices, and ultimately with one's sense

of personhood. It is difficult to see this as at all controversial. But since Part IV will eventually come to question it, we had better observe right away that it is no more than an assumption, subject in principle to empirical examination.

However, that is not the issue I want to take up at the moment. At stake is a much more basic clarification of the relation of the phenomenological mind to the computational mind.

2.3 The Mind-Mind Problem

The problem is this: just because the computational mind and the phenomenological mind are both different domains of description from the physical body, this does not mean they are the *same* domain.

Optimists on this issue (such as Hofstadter (1979), if I read him properly) say that if a nervous system (or a computer) can just achieve some sufficient degree of complexity, consciousness will somehow miraculously emerge. In fact, though, I find it every bit as incoherent to speak of conscious experience as a flow of information as to speak of it as a collection of neural firings. It is completely unclear to me how computations, no matter how complex or abstract, can add up to an experience.

Hofstadter's position, like some others, arises in part from confusing consciousness with self-consciousness (that is, reflective awareness). The latter involves (at least) the combination of ordinary consciousness with self-reference. Self-reference, of course, is a capacity common to people, certain kinds of computer languages, and ordinary sentences such as "The sentence I am now uttering is ten words long"; it has nothing to do with consciousness per se. Hofstadter, preoccupied with self-reference, neglects the other essential components of self-awareness. Then, finding both recursive self-reference and consciousness mind-boggling (and even inspiring of religious awe!), he uncritically identifies the former as the source of the latter.

More sober consideration suggests that the leap from self-reference to self-consciousness to consciousness is unwarranted. Consider the problem of qualia. As Block (1978) points out, no computational theory gives the slightest idea of how to get blueness or saltiness or painfulness out of computations. Like neurological accounts, computational accounts may provide the right *distinctions*—they may, as it were, give the phenomenological mind the cue to produce experiences of blue at the right times and experiences of red at the right times. But that is not the same as producing the experiences themselves.

With the problem of form there is little more hope. Certainly, "propositional" representations of the general form "X is square" (including semantic networks as well as language-like representations—see section

8.1) provide no basis for the *experience* of squareness, whatever their virtues in other respects. Representations more geometric in character have been proposed to account for visual imagery (Shepard and Cooper 1982; Kosslyn 1980; see section 9.5), and these seem at first blush more satisfactory. But, as we will see, the force behind these proposals is only that such representations can encode the proper distinctions among geometric forms, not that they can account for the experience of forms per se. It is only by slipping into thinking of them as "pictures in the head," viewed by the "mind's eye," that we trick ourselves into believing otherwise.

Finally, the externalization of experience creates the same problems for the computational mind as it does for the neurological brain. I experience objects out there *in the world*, and pain *in my toe*, not computational states in my brain. The computational mind may well express the distinctions among the locations of experienced objects, but it is hard to see how it can literally put them there.

A curiously distorted version of these points appears in Searle's (1980) widely cited attempt to refute the computational theory of mind. The argument centers on *intentionality*, which Searle defines as "that feature of certain mental states by which they are directed at or about objects and states of affairs in the world" (p. 424). He argues correctly that all a computer can do is manipulate its own internal symbols; it cannot understand those symbols as connected to an external reality. From the point of view of the machine the symbols are not symbols *for anything*—they are just meaningless marks. Thus, the computer's states are not intentional in Searle's sense. He concludes that "the brain's casual capacity to produce intentionality cannot consist in its instantiating a computer program, since for any program you like it is possible for something to instantiate that program and still not have any [intentional] mental states. Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality" (p. 424).

If we strip away the jargon, I think Searle's argument can be seen as a displacement of a more fundamental and gut-level claim: that running a computer program does not produce consciousness. What makes us believe that our mental states are intentional in Searle's sense is that we experience the things in the world that our thoughts are about. What makes us believe that a computer's states are not intentional is that we can't imagine how the computer could experience the world. In short, Searle's peculiar "causal capacity of the brain to produce intentionality," which is lacking in computers and in computational theories, seems to be a euphemism for conscious awareness.

A survey of the varieties of experience suggests in fact that intentionality in Searle's sense is the wrong place to locate the problem. If intention-

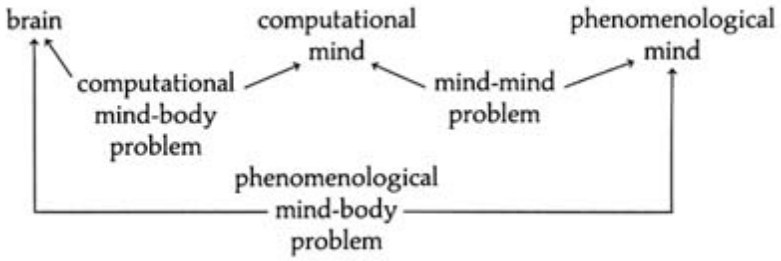

Figure 2.1
Domains of psychology and their relations

ality concerns the relation between mental states and the (real) world, it leaves no room for the "aboutness" of mental states that produce images or hallucinations. That is, a theory that focuses on the intentionality of mental states suffers from the same difficulty as the theories of mind that involve "grasping" external reality, as discussed in section 1.3. A more inclusive formulation might take intentionality to be the property of (computational) mental states whereby they are related to the world as experienced, whether real or not. But then it boils down again to the issue of consciousness. (I return to this and intimately related issues in sections 7.4. and 7.5.)

In this reading of Searle, then, I take him to be making (albeit indirectly) the same point I have been pressing here: the computational mind offers no explication of what a conscious experience is. This is not necessarily reason to reject the computational mind, as Searle does. But it is reason to limit the claims one makes about the power of the computer analogy.

The upshot is that psychology now has not two domains to worry about, brain and mind, but three: the brain, the computational mind, and the phenomenological mind. Consequently, Descartes's formulation of the mind-body problem is split into two separate issues. The "phenomenological mind-body problem" (our concern in chapter 1) is, How can a brain have experiences? The "computational mind-body problem" is, How can a brain accomplish reasoning? In addition, we have the mind-*mind* problem, namely, What is the relationship between computational states and experience? Figure 2.1 sketches these relationships.

From this vantage let us return to Putnam's (1960) claim that seeing the brain as an information processor solves the mind-body problem. What Putnam has actually done is to discover that functionalism provides an *approach* to the computational mind-body problem; he neglects the phenomenological mind-body problem altogether. A *solution* to the computational mind-body problem would tell us how the information structures and processes of the computational mind are neurally instantiated, that is, how the software runs on the hardware. Since at the moment we

know relatively little about either the software or the hardware in the requisite fine detail (for example, how the neurons accomplish visual identification or sentence comprehension), Putnam's claim of a solution would seem somewhat premature.

However, we do understand the problem fairly well in principle. We can hope for at least what Fodor (1975) has called *token reductionism*. A description of this sort would be able to say that such-and-such a computational structure or process taking place in such-and-such an individual on such-and-such an occasion is instantiated by such-and-such neural structures or processes in that individual's nervous system. On the other hand, it does not seem reasonable to hope for what Fodor calls *type reductionism*, in which one could say, for instance, that the concept *grandmother* is instantiated in everyone in exactly the same way. A position somewhere between these two is probably the best we can expect: token reduction for many kinds of structures and processes, type reduction for others (for instance, the low-level visual system), and beyond this some taxonomy of types (for instance, that such-and-such types of information and processes are localized in such-and-such an area of the brain and performed by such-and-such arrangements of neural architecture).

Whatever the ultimate outcome on these issues, the overall form of the solution to the computational mind-body problem is clear. Just as we say that a computer program is a way of specifying the operation of the machine in terms of its functional organization, so we can regard the computational mind as an abstract specification of functional organization in the nervous system—even if, at the moment, we cannot translate from this description into hardware terms. Hence, if one were to choose which theory of the mind-body relationship applies to this case, the appropriate answer would clearly be an identity theory: the computational mind is another way of describing the brain. It would make little sense to adopt a "behaviorist" position prohibiting discourse about information structure; it would make even less sense to adopt an "epiphenomenalist" position in which somehow the information processing went on in a metaphysically distinct domain.

2.4 Positions on the Mind-Mind Problem; Theory II

Such clarity as there is in this relationship vanishes, it seems to me, when we turn to the relation of the computational mind to the phenomenological mind. Rather, we are in the same situation as we were with the phenomenological mind-body problem. Although the computational theory of mind may be of help in elucidating the units and distinctions that are present to experience, and although the organization of the phenomenological mind may be more closely paralleled by the computational mind than by raw