1 MEAT MACHINES
Mindware as Software

1.1 Sketches

The computer scientist Marvin Minsky once described the human brain as a meat machine—no more no less. It is, to be sure, an ugly phrase. But it is also a striking image, a compact expression of both the genuine scientific excitement and the rather gung-ho materialism that tended to characterize the early years of cognitive scientific research. Mindware—our thoughts, feelings, hopes, fears, beliefs, and intellect—is cast as nothing but the operation of the biological brain, the meat machine in our head. This notion of the brain as a meat *machine* is interesting, for it immediately invites us to focus not so much on the material (the meat) as on the machine: the way the material is organized and the kinds of operation it supports. The same machine (see Box 1.1) can, after all, often be made of iron, or steel, or tungsten, or whatever. What we confront is thus both a rejection of the idea of mind as immaterial spirit-stuff and an affirmation that mind is best studied from a kind of engineering perspective that reveals the nature of the machine that all that wet, white, gray, and sticky stuff happens to build.

What exactly is meant by casting the brain as a machine, albeit one made out of meat? There exists a historical trend, to be sure, of trying to understand the workings of the brain by analogy with various currently fashionable technologies: the telegraph, the steam engine, and the telephone switchboard are all said to have had their day in the sun. But the "meat machine" phrase is intended, it should now be clear, to do more than hint at some rough analogy. For with regard to the very special class of machines known as computers, the claim is that the brain (and, by

Box 1.1

THE "SAME MACHINE"

In what sense can "the same machine" be made out of iron, or steel, or whatever? Not, obviously, in the strict sense of numerical identity. A set of steel darts and a set of tungsten ones cannot be the *very same* (numerically identical) set of darts. The relevant sense of sameness is, rather, some sense of *functional* sameness. You can make a perfectly *good* set of darts out of either material (though not, I suppose, out of jello), just as you can make a *perfectly* good corkscrew using a myriad (in this latter case quite radically) different designs and materials. In fact, what *makes* something a corkscrew is simply that it is designed as, and is capable of acting as, a cork-removing device. The notion of a brain as a meat machine is meant to embody a similar idea: that what matters about the brain is not the stuff it is made of but the way that stuff is organized so as to support thoughts and actions. The idea is that this capability depends on quite abstract properties of the physical device that could very well be duplicated in a device made, say, out of wires and silicon. Sensible versions of this idea need not claim then that *any* material will do: perhaps, for example, a certain stability over time (a tendency not to rapidly disorganize) is needed. The point is just that given that certain preconditions are met the same functionality can be pressed from multiple different materials and designs. For some famous opposition to this view, see Searle (1980, 1992).

not unproblematic extension, the mind) actually *is* some such device. It is not that the brain is somehow *like* a computer: everything is like everything else in some respect or other. It is that neural tissues, synapses, cell assemblies, and all the rest are just nature's rather wet and sticky way of building a hunk of honest-to-God computing machinery. Mindware, it is then claimed, is found "in" the brain in just the way that software is found "in" the computing system that is running it.

The attractions of such a view can hardly be overstated. It makes the mental special without making it ghostly. It makes the mental depend on the physical, but in a rather complex and (as we shall see) liberating way. And it provides a ready-made answer to a profound puzzle: how to get sensible, reason-respecting behavior out of a hunk of physical matter. To flesh out this idea of nonmysterious reason-respecting behavior, we next review some crucial developments[1] in the history (and prehistory) of artificial intelligence.

[1]The next few paragraphs draw on Newell and Simon's (1976) discussion of the development of the Physical Symbol Hypothesis (see Chapter 2 following), on John Haugeland's (1981a), and on Glymour, Ford, and Hayes' (1995).

One key development was the appreciation of the power and scope of formal logics. A decent historical account of this development would take us too far afield, touching perhaps on the pioneering efforts in the seventeenth century by Pascal and Leibniz, as well as on the twentieth-century contributions of Boole, Frege, Russell, Whitehead, and others. A useful historical account can be found in Glymour, Ford, and Hayes (1995). The idea that shines through the history, however, is the idea of finding and describing "laws of reason"—an idea whose clearest expression emerged first in the arena of formal logics. Formal logics are systems comprising sets of symbols, ways of joining the symbols so as to express complex propositions, and rules specifying how to legally derive new symbol complexes from old ones. The beauty of formal logics is that the steadfast application of the rules guarantees that you will never legally infer a false conclusion from true premises, even if you have no idea what, if anything, the strings of symbols actually mean. Just follow the rules and truth will be preserved. The situation is thus a little (just a little) like a person, incompetent in practical matters, who is nonetheless able to successfully build a cabinet or bookshelf by following written instructions for the manipulation of a set of preprovided pieces. Such building behavior can look as if it is rooted in a deep appreciation of the principles and laws of woodworking: but in fact, the person is just blindly making the moves allowed or dictated by the instruction set.

Formal logics show us how to preserve at least one kind of semantic (meaning-involving: see Box 1.2) property without relying on anyone's actually appreciating the meanings (if any) of the symbol strings involved. The seemingly ghostly and ephemeral world of meanings and logical implications is respected, and in a certain sense recreated, in a realm whose operating procedures do not rely on meanings at all! It is recreated as a realm of marks or "tokens," recognized by their physical ("syntactic") characteristics alone and manipulated according to rules that refer only to those physical characteristics (characteristics such as the shape of the symbol—see Box 1.2). As Newell and Simon comment:

> Logic . . . was a game played with meaningless tokens according to certain purely syntactic rules. Thus progress was first made by walking away from all that seemed relevant to meaning and human symbols. (Newell and Simon, 1976, p. 43)

Or, to put it in the more famous words of the philosopher John Haugeland:

> If you take care of the syntax, *the semantics will take care of itself.* (Haugeland, 1981a, p. 23, original emphasis)

This shift from meaning to form (from semantics to syntax if you will) also begins to suggest an attractive liberalism concerning actual physical structure. For what matters, as far as the identity of these formal systems is concerned, is not, e.g., the precise shape of the symbol for "and." The shape could be "AND" or "and" or "&" or "$\wedge$" or whatever. All that matters is that the shape is used consistently and that the rules are set up so as to specify how to treat strings of symbols joined by that shape: to allow, for example, the derivation of "A" from the string "A and

Box 1.2

SYNTAX AND SEMANTICS

Semantic properties are the "meaning-involving" properties of words, sentences, and internal representations. *Syntactic* properties, at least as philosophers tend to use the term, are nonsemantic properties of, e.g., written or spoken words, or of any kinds of inscriptions of meaningful items (e.g., the physical states that the pocket calculator uses to store a number in memory). Two synonymous written words ("dog" and "chien") are thus semantically identical but syntactically distinct, whereas ambiguous words ("bank" as in river or "bank" as in high street) are syntactically identical but semantically distinct. The idea of a *token* is the idea of a specific syntactic item (e.g., *this* occurrence of the word "dog"). A pocket calculator manipulates physical tokens (inner syntactic states) to which the operation of the device is sensitive. It is by being sensitive to the distinct syntactic features of the inner tokens that the calculator manages to behave in an arithmetic-respecting fashion: it is set up *precisely* so that syntax-driven operations on inner tokens standing for numbers respect meaningful arithmetical relations between the numbers. Taking care of the syntax, in Haugeland's famous phrase, thus allows the semantics to take care of itself.

B." Logics are thus first-rate examples of *formal systems* in the sense of Haugeland (1981a, 1997). They are systems whose essence lies not in the precise physical details but in the web of legal moves and transitions.

Most games, Haugeland notes, are formal systems in exactly this sense. You can play chess on a board of wood or marble, using pieces shaped like animals, movie stars, or the crew of the star ship Enterprise. You could even, Haugeland suggests, play chess using helicopters as pieces and a grid of helipads on top of tall buildings as the board. All that matters is again the web of legal moves and the physical distinguishability of the tokens.

Thinking about formal systems thus liberates us in two very powerful ways at a single stroke. Semantic relations (such as truth preservation: if "A and B" is true, "A" is true) are seen to be respected in virtue of procedures that make no intrinsic reference to meanings. And the specific physical details of any such systems are seen to be unimportant, since what matters is the golden web of moves and transitions. Semantics is thus made unmysterious without making it brute physical. Who says you can't have your cake and eat it?

The next big development was the formalization (Turing, 1936) of the notion of computation itself. Turing's work, which predates the development of the digital computer, introduced the foundational notion of (what has since come to be known as) the Turing machine. This is an imaginary device consisting of an infinite tape, a simple processor (a "finite state machine"), and a read/write head. The tape acts as data store, using some fixed set of symbols. The read/write head can read a symbol off the tape, move itself one square backward or forward on the tape, and write onto the tape. The finite state machine (a kind of central processor) has enough memory to recall what symbol was just read and what state it (the finite state machine) was in. These two facts together determine the next action, which is carried out by the read/write head, and determine also the next state of the finite state machine. What Turing showed was that some such device, performing a sequence of simple computations governed by the symbols on the tape, could compute the answer to any sufficiently well-specified problem (see Box 1.3).

We thus confront a quite marvelous confluence of ideas. Turing's work clearly suggested the notion of a physical machine whose syntax-following properties would enable it to solve any well-specified problem. Set alongside the earlier work on logics and formal systems, this amounted to nothing less than

> . . . the emergence of a new level of analysis, independent of physics yet mechanistic in spirit . . . a science of structure and function divorced from material substance. (Pylyshyn, 1986, p. 68)

Thus was classical cognitive science conceived. The vision finally became flesh, however, only because of a third (and final) innovation: the actual construction of general purpose electronic computing machinery and the development of flexible, high-level programming techniques. The bedrock machinery (the digital computer) was designed by John von Neumann in the 1940s and with its advent all the pieces seemed to fall finally into place. For it was now clear that once realized in the physical medium of an electronic computer, a formal system could run *on its own*, without a human being sitting there deciding how and when to apply the rules to initiate the legal transformations. The well-programmed electronic computer, as John Haugeland nicely points out, is really just an automatic ("self-moving") formal system:

> It is like a chess set that sits there and plays chess by itself, without any intervention from the players, or an automatic formal system that writes out its own proofs and theorems without any help from the mathematician. (Haugeland, 1981a, p. 10; also Haugeland, 1997, pp. 11–12)

Of course, the machine needs a program. And programs were, in those days (but see Chapter 4), written by good old-fashioned human beings. But once the program was in place, and the power on, the machine took care of the rest. The transitions between legal syntactic states (states that also, under interpretation, *meant* something) no longer required a human operator. The physical world suddenly included clear, nonevolved, nonorganic examples of what Daniel Dennett would later dub "syntactic engines"—quasiautonomous systems whose sheer physical make-

Box 1.3

A Turing Machine

To make the idea of Turing machine computation concrete, let us borrow an example from Kim (1996, pp. 80–85). Suppose the goal is to get a Turing machine to add positive numbers. Express the numbers to be added as a sequence of the symbols "#" (marking the beginning and end of numbers) "1" and "+." So the sum 3 + 2 is encoded on the tape as shown in Figure 1.1. A neat program for adding the numbers (where "Λ A" indicates the initial location and initial state of the read/write head) is as follows:

Instruction 1: If read-write head is in machine state A and encounters a "1," it moves one square to the right, and the head stays in state A.

Instruction 2: If the head is in state A and encounters a "+," it replaces it with a "1," stays in state A, and moves one square to the right.

Instruction 3: If the head is in state A and it encounters a "#," move one square left and go into machine state B.

Instruction 4: If the head is in machine state B and encounters a "1," delete it, replace with a "#," and halt.

You should be able to see how this works. Basically, the machine starts "pointed" at the leftmost "1." It scans right seeking a "+," which it replaces with a "1." It continues scanning right until the "#" indicates the end of the sum, at which point it moves one square left, deletes a single "1," and replaces it with a "#." The tape now displays the answer to the addition problem in the same notation used to encode the question, as shown in Figure 1.2.

Similar set-ups (try to imagine how they work) can do subtraction, multiplication, and more (see Kim, 1996, pp. 83–85). But Turing's most strik-

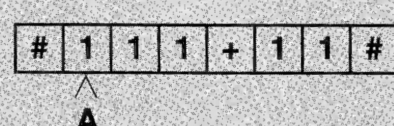

Figure 1.1 (After Kim, 1996, p. 81.)

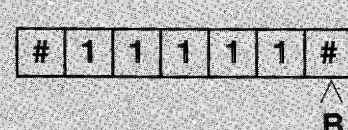

Figure 1.2 (After Kim, 1996, p. 81.)

ing achievement in this area was to show that you could then define a special kind of Turing machine (the aptly-named universal Turing machine) able to imitate any other Turing machine. The symbols on the tape, in this universal case, encode a description of the behavior of the other machine. The universal Turing machine uses this description to mimic the input–output function of any other such device and hence is itself capable of carrying out *any* sufficiently well-specified computation. (For detailed accounts see Franklin, 1995; Haugeland, 1985; Turing, 1936, 1950.)

The Turing machine affords a fine example of a simple case in which syntax-driven operations support a semantics-respecting (meaning-respecting) process. Notice also that you could *build* a simple Turing machine out of many different materials. It is the formal (syntactic) organization that matters for its semantic success.

up ensured (under interpretation) some kind of ongoing reason-respecting behavior. No wonder the early researchers were jubilant! Newell and Simon nicely capture the mood:

> It is not my aim to surprise or shock you. . . . But the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be co-extensive with the range to which the human mind has been applied. (Newell and Simon, 1958, p. 6, quoted in Dreyfus and Dreyfus, 1990, p. 312)

This jubilant mood deepened as advanced programming techniques[2] brought forth impressive problem-solving displays, while the broader theoretical and philosophical implications (see Box 1.4) of these early successes could hardly have been more striking. The once-mysterious realm of mindware (represented, admittedly, by just two of its many denizens: truth preservation and abstract problem solving) looked ripe for conquest and understanding. Mind was not ghostly stuff, but the operation of a formal, computational system implemented in the meatware of the brain.

Such is the heart of the matter. Mindware, it was claimed, is to the neural meat machine as software is to the computer. The brain may be the standard (local, earthly, biological) implementation—but cognition is a program-level thing. Mind

[2] For example, list-processing languages, as pioneered in Newell and Simon's Logic Theorist program in 1956 and perfected in McCarthy's LISP around 1960, encouraged the use of more complex "recursive programming" strategies in which symbols point to data structures that contain symbols pointing to further data structures and so on. They also made full use of the fact that the same electronic memory could store both program and data, a feature that allowed programs to be modified and operated on in the same ways as data. LISP even boasted a universal function, EVAL, that made it as powerful, modulo finite memory limitations, as a Universal Turing Machine.

Box 1.4

MACHINE FUNCTIONALISM

The leading philosophical offspring of the developments in artificial intelligence went by the name of machine functionalism, and it was offered as an answer to one of the deepest questions ever asked by humankind, viz. what is the essence (the deep nature) of the mental? What fundamental facts make it the case that some parts of the physical world have mental lives (thoughts, beliefs, feelings, and all the rest) and others do not? Substance dualists, recall, thought that the answer lay in the presence or absence of a special kind of mental *stuff*. Reacting against this idea (and against so-called philosophical behaviorism—see Appendix I). Mind–brain identity theorists, such as Smart (1959) (and again, see Appendix I), claimed that mental states *just are* processes going on in the brain. This bald identity claim, however, threatened to make the link between mental states and specific, material brain states a little too intimate. A key worry (e.g., Putnam, 1960, 1967) was that if it was really essential to being in a certain mental state that one be in a specific brain state, it would seem to follow that creatures lacking brains built just like ours (say, Martians or silicon-based robots) could not be in those very same mental states. But surely, the intuition went, creatures with very different brains from ours could, at least in principle, share, e.g., the belief that it is raining. Where, then, should we look for the commonality that could unite the robot, the Martian, and the Bostonian? The work in logic and formal systems, Turing machines, and electronic computation now suggested an answer: look not to the specific physical story (of neurons and wetware), nor to the surface behavior, but to the inner organization, that is to say, to the golden web: to the abstract, formal organization of the system. It is this organization—depicted by the machine functionalists as a web of links between possible inputs, inner computational states, and outputs (actions, speech)—that fixes the shape and contents of a mental life. The building materials do not matter: the web of transitions could be realized in flesh, silicon, or cream cheese (Putnam, 1975, p. 291). To be in such and such a mental state is simply to be a physical device, of whatever composition, that satisfies a specific formal description. Mindware, in humans, happens to run on a meat machine. But the very same mindware (as picked out by the web of legal state transitions) might run in some silicon device, or in the alien organic matter of a Martian.

is thus ghostly enough to float fairly free of the gory neuroscientific details. But it is not so ghostly as to escape the nets of more abstract (formal, computational) scientific investigation. This is an appealing story. But is it correct? Let's worry.

1.2 Discussion

(A brief note of reassurance: many of the topics treated below recur again and again in subsequent chapters. At this point, we lack much of the detailed background needed to really do them justice. But it is time to test the waters.)

A. WHY TREAT THOUGHT AS COMPUTATION?

Why treat thought as computation? The principal reason (apart from the fact that it seems to work!) is that thinkers are physical devices whose behavior patterns are reason respecting. Thinkers act in ways that are usefully understood as sensitively guided by reasons, ideas, and beliefs. Electronic computing devices show us one way in which this strange "dual profile" (of physical substance and reason-respecting behavior) can actually come about.

The notion of reason-respecting behavior, however, bears immediate amplification. A nice example of this kind of behavior is given by Zenon Pylyshyn. Pylyshyn (1986) describes the case of the pedestrian who witnesses a car crash, runs to a telephone, and punches out 911. We could, as Pylyshyn notes, try to explain this behavior by telling a purely physical story (maybe involving specific neurons, or even quantum events, whatever). But such a story, Pylyshyn argues, will not help us understand the behavior in its *reason-guided* aspects. For example, suppose we ask: what would happen if the phone was dead, or if it was a dial phone instead of a touch-tone phone, or if the accident occurred in England instead of the United States? The neural story underlying the behavioral response will differ widely if the agent dials 999 (the emergency code in England) and not 911, or must run to find a working phone. Yet common sense psychological talk makes sense of all these options at a stroke by depicting the agent as seeing a crash and *wanting to get help*. What we need, Pylyshyn powerfully suggests, is a scientific story that remains in touch with this more abstract and reason-involving characterization. And the simplest way to provide one is to imagine that the agent's brain contains states ("symbols") that represent the event *as* a car crash and that the computational state-transitions occurring inside the system (realized as physical events in the brain) then lead to new sets of states (more symbols) whose proper interpretation is, e.g., "seek help," "find a telephone," and so on. The interpretations thus glue inner states to sensible real-world behaviors. Cognizers, it is claimed, "instantiate . . . representation physically as cognitive codes and . . . their behavior is a causal consequence of operations carried out on those codes" (Pylyshyn, 1986, p. xiii).

The same argument can be found in, e.g., Fodor (1987), couched as a point about content-determined transitions in trains of thought, as when the thought "it is raining" leads to the thought "let's go indoors." This, for Fodor (but see Chapters 4 onward), is the essence of human rationality. How is such rationality mechanically possible? A good empirical hypothesis, Fodor suggests, is that there are neural symbols (inner states apt for interpretation) that mean, e.g., "it is raining" and whose physical properties lead in context to the generation of other symbols that mean "let's go indoors." If that is how the brain works then the brain is indeed a computer in exactly the sense displayed earlier. And if such were the case, then the mystery concerning reason-guided (content-determined) transitions in thought is resolved:

> If the mind is a sort of computer, we begin to see how . . . there could be non-arbitrary content-relations among causally related thoughts. (Fodor, 1987, p. 19)

Such arguments aim to show that the mind *must* be understood as a kind of computer implemented in the wetware of the brain, on pain of failing empirically to account for rational transitions among thoughts. Reason-guided action, it seems, makes good scientific sense if we imagine a neural economy organized as a syntax-driven engine that tracks the shape of semantic space (see, e.g., Fodor, 1987, pp. 19–20).

B. IS SOFTWARE AN AUTONOMOUS LEVEL IN NATURE?

The mindware/software equation is as beguiling as it is, at times, distortive. One immediate concern is that all this emphasis on algorithms, symbols, and programs tends to promote a somewhat misleading vision of *crisp level distinctions in nature.* The impact of the theoretical independence of algorithms from hardware is an artifact of the long-term neglect of issues concerning real-world action taking and the time course of computations. For an algorithm or program as such is just a sequence of steps with no inbuilt relation to real-world timing. Such timing depends crucially on the particular way in which the algorithm is implemented on a real device. Given this basic fact, the theoretical independence of algorithm from hardware is unlikely to have made much of an impact on Nature. We must expect to find biological computational strategies closely tailored to getting useful real-time results from available, slow, wetware components. In practice, it is thus unlikely that we will be able to fully appreciate the formal organization of natural systems without some quite detailed reference to the nature of the neural hardware that provides the supporting implementation. In general, attention to the nature of real biological hardware looks likely to provide both important clues about and constraints on the kinds of computational strategy used by real brains. This topic is explored in more depth in Chapters 4 through 6.

Furthermore, the claim that mindware is software is—to say the least—merely schematic. For the space of possible types of explanatory story, all broadly computational (but see Box 1.5), is very large indeed. The comments by Fodor and by

Box 1.5

What Is Computation?

It is perhaps worth mentioning that the foundational notion of computation is itself still surprisingly ill understood. What do we really mean by calling some phenomenon "computational" in the first place? There is no current consensus at least (in the cognitive scientific community) concerning the answer to this question. It is mostly a case of "we know one when we see one." Nonetheless, there is a reasonable consensus concerning what I'll dub the "basic profile," which is well expressed by the following statement:

> we count something as a computer because, and only when, its inputs and outputs can be usefully and systematically interpreted as representing the ordered pairs of some function that interests us. (Churchland and Sejnowski, 1992, p. 65)

Thus consider a pocket calculator. This physical device computes, on this account, because first, there is a reliable and systematic way of interpreting various states of the device (the marks and numerals on the screen and keyboard) as representing other things (numbers). And second, because the device is set up so that under that interpretation, its physical state changes mirror semantic (meaningful) transitions in the arithmetical domain. Its physical structure thus forces it to respect mathematical constraints so that inputs such as "4 × 3" lead to outputs such as "12" and so on.

A truly robust notion of the conditions under which some actual phenomenon counts as computational would require, however, some rather more *objective* criterion for determining when an encountered (nondesigned) physical process is actually implementing a computation—some criterion that does not place our interpretive activities and interests so firmly at center stage.

The best such account I know of is due to Dave Chalmers (1996, Chapter 9). Chalmers' goal is to give an "objective criterion for implementing a computation" (p. 319). Intuitively, a physical device 'implements' an abstract, formal computational specification just in case the physical device is set up to undergo state changes that march in step with those detailed in the specification. In this sense a specific word-processing program might, for example, constitute a formal specification that can (appropriately configured) be made to run on various kinds of physical device (MACS, PCs, etc.).

Chalmers' proposal, in essence, is that a physical device implements an abstract formal description (a specification of states and state-transition relations) just in case "the causal structure of the system mirrors the formal structure of the computation" (1996, p. 317). The notion of mirroring is then cashed out in terms of a fairly fine-grained mapping of states and state changes in the physical device onto the elements and transitions present in the abstract specification. Chalmer's allows that every physical system will implement some computational description. But the appeal to fine-grained mappings is meant to ensure that you cannot interpret *every* physical system as implementing *every* computational description. So although the claim that the brain implements *some* computational description is fairly trivial, the claim that it implements a *specific* computational description is not. And it is the brain's implementation of a specific computational description that is meant to explain mental properties.

The computational profile of most familiar devices is, of course, the result of the deliberate imposition of a mapping, via some process of intelligent design. But the account is not intrinsically so restricted. Thus suppose some creature has evolved organic inner states that represent matters of adaptive importance such as the size, number, and speed of approach of predators. If that evolutionary process results in a physical system whose causal state transitions, under that interpretation, make semantic sense (e.g., if fewer than two predators detected cause a "stand and fight" inner token leading to aggressive output behavior, whereas three or more yield a "run and hide" response), then Nature has, on this account, evolved a small computer. The brain, if the conjectures scouted earlier prove correct, is just such a natural computer, incorporating inner states that represent external events (such as the presence of predators) and exploiting state-transition routines that make sensible use of the information thus encoded.

Pylyshyn do, it is true, suggest a rather specific kind of computational story (one pursued in detail in the next chapter). But the bare explanatory schema, in which semantic patterns emerge from an underlying syntactic, computational organization, covers a staggeringly wide range of cases. The range includes, for example, standard artificial intelligence (A.I.) approaches involving symbols and rules, "connectionist" approaches that mimic something of the behavior of neural assemblies (see Chapter 4), and even Heath Robinsonesque devices involving liquids, pulleys, and analog computations. Taken very liberally, the commitment to understanding mind as the operation of a syntactic engine can amount to little more than a bare assertion of physicalism—the denial of spirit-stuff.[3]

To make matters worse, a variety of different computational stories may be told about one and the same physical device. Depending on the grain of analysis

[3]Given our notion of computation (see Box 1.5), the claim is just a little stronger, since it also requires the presence of systematically interpretable inner states, i.e., internal representations.

used, a single device may be depicted as carrying out a complex parallel search or as serially transforming an input *x* into an output *y*. Clearly, what grain we choose will be determined by what questions we hope to answer. Seeing the transition as involving a nested episode of parallel search may help explain specific error profiles or why certain problems take longer to solve than others, yet treating the process as a simple unstructured transformation of *x* to *y* may be the best choice for understanding the larger scale organization of the system. There will thus be a constant interaction between our choice of explanatory targets and our choice of grain and level of computational description. In general, there seems little reason to expect a single type or level of description to do all the work we require. Explaining the relative speed at which we solve different problems, and the kinds of interference effects we experience when trying to solve several problems at once (e.g., remembering two closely similar telephone numbers), may well require explanations that involve very specific details about how inner representations are stored and structured, whereas merely accounting for, e.g., the bare facts about rational transitions between content-related thoughts may require only a coarser grained computational gloss. [It is for precisely this reason that connectionists (see Chapter 4) describe themselves as exploring the microstructure of cognition.] The explanatory aspirations of psychology and cognitive science, it seems clear, are sufficiently wide and various as to require the provision of explanations at a variety of different levels of grain and type.

In sum, the image of mindware as software gains its most fundamental appeal from the need to accommodate reason-guided transitions in a world of merely physical flux. At the most schematic level, this equation of mindware and software is useful and revealing. But we should not be misled into believing either (1) that "software" names a single, clearly understood level of neural organization or (2) that the equation of mindware and software provides any deep warrant for cognitive science to ignore facts about the biological brain.

C. MIMICKING, MODELING, AND BEHAVIOR

Computer programs, it often seems, offer only shallow and brittle simulacrums of the kind of understanding that humans (and other animals) manage to display. Are these just teething troubles, or do the repeated shortfalls indicate some fundamental problem with the computational approach itself? The worry is a good one. There are, alas, all too many ways in which a given computer program may merely mimic, but not illuminate, various aspects of our mental life. There is, for example, a symbolic A.I. program that does a very fine job of mimicking the verbal responses of a paranoid schizophrenic. The program ("PARRY," Colby, 1975; Boden, 1977, Chapter 5) uses tricks such as scanning input sentences for key words (such as "mother") and responding with canned, defensive outbursts. It is capable, at times, of fooling experienced psychoanalysts. But no one would claim that it is a useful psychological model of paranoid schizophrenia, still less that it is (when up and running on a computer) a paranoid schizophrenic itself!

Or consider a chess computer such as Deep Blue. Deep Blue, although capable of outstanding play, relies heavily on the brute-force technique of using its superfast computing resources to examine all potential outcomes for up to seven moves ahead. This strategy differs markedly from that of human grandmasters, who seem to rely much more on stored knowledge and skilled pattern recognition (see Chapter 4). Yet, viewed from a certain height, Deep Blue is not a bad simulation of human chess competence. Deep Blue and the human grandmaster are, after all, more likely to agree on a particular move (as a response to a given board state) than are the human grandmaster and the human novice! At the level of gross input–output profiles, the human grandmaster and Deep Blue are thus clearly similar (not identical, as the difference in underlying strategy—brute force versus pattern recognition—sometimes shines through). Yet once again, it is hard to avoid the impression that all that the machine is achieving is top-level mimicking: that there is something amiss with the underlying strategy that either renders it unfit as a substrate for a real intelligence, or else reveals it as a kind of intelligence very alien to our own.

This last caveat is important. For we must be careful to distinguish the question of whether such and such a program constitutes a good model of *human* intelligence from the question of whether the program (when up and running) displays some kind of *real, but perhaps nonhuman* form of intelligence and understanding. PARRY and Deep Blue, one feels, fail on both counts. Clearly, neither constitutes a faithful psychological model of the inner states that underlie human performance. And something about the basic style of these two computational solutions (canned sentences activated by key words, and brute-force look-ahead) even makes us uneasy with the (otherwise charitable) thought that they might nonetheless display real, albeit alien, kinds of intelligence and awareness.

How, though, are we to decide what kinds of computational substructure *might* be appropriate? Lacking, as we must, first-person knowledge of what (if anything) it is like to be PARRY or Deep Blue, we have only a few options. We could insist that all real thinkers must solve problems using exactly the same kinds of computational strategy as human brains (too anthropocentric, surely). We could hope, optimistically, for some future scientific understanding of the *fundamentals* of cognition that will allow us to recognize (on broad theoretical grounds) the shape of alternative, but genuine, ways in which various computational organizations might support cognition. Or we could look to the gross behavior of the systems in question, insisting, for example, on a broad and flexible range of responses to a multiplicity of environmental demands and situations. Deep Blue and PARRY would then fail to make the grade not merely because their inner organizations looked alien to us (an ethically dangerous move) but because the behavioral repertoire they support is too limited. Deep Blue cannot recognize a mate (well, only a check-

...that it is (when

... although capa-
...of using its su-
...for up to seven
...grandmasters,
...pattern recognition
...not a bad simula-
...grandmaster are, af-
...to a given board
...At the level of gross
...are thus clearly sim-
...brute force versus pat-
...it is hard to avoid
...level mimicking: that
...either renders it unfit
...kind of intelligence very

...to distinguish the ques-
...a good model of *human*
...when up and running)
...of intelligence and under-
...counts. Clearly, neither
...that underlie human
...two computational so-
...force look-ahead) even
...that they might nonethe-
...awareness.
...onal substructure *might*
...of what (if anything)
...options. We could insist
...the same kinds of compu-
...surely). We could hope,
...the *fundamentals* of cog-
...grounds) the shape of
...onal organizations might
...of the systems in ques-
...of responses to a multi-
...Blue and PARRY would
...organizations looked
...the behavioral repertoire
...a mate (well, only a check-

mate!), nor cook an omelette. PARRY cannot decide to become a hermit or take up the harmonica, and so on.

This move to behavior is not without its own problems and dangers, as we will see in Chapter 3. But it should now be clearer why some influential theorists (especially Turing, 1950) argued that a sufficient degree of behavioral success should be allowed to settle the issue and to establish once and for all that a candidate system is a genuine thinker (albeit one whose inner workings may differ greatly from our own). Turing proposed a test (now known as the Turing Test) that involved a human interrogator trying to spot (from verbal responses) whether a hidden conversant was a human or a machine. Any system capable of fooling the interrogator in ongoing, open-ended conversation, Turing proposed, should be counted as an intelligent agent. Sustained, top-level verbal behavior, if this is right, is a sufficient test for the presence of real intelligence. The Turing Test invites consideration of a wealth of issues that we cannot dwell on here (several surface in Chapter 3). It may be, for example, that Turing's original restriction to a verbal test leaves too much scope for "tricks and cheats" and that a better test would focus more heavily on real-world activity (see Harnad, 1994).

It thus remains unclear whether we should allow that surface behaviors (however complex) are sufficient to distinguish (beyond all theoretical doubt) real thinking from mere mimicry. Practically speaking, however, it seems less morally dangerous to allow behavioral profiles to lead the way (imagine that it is discovered that you and you alone have a mutant brain that uses brute-force, Deep Blue-like strategies where others use quite different techniques: has science discovered that *you* are not a conscious, thinking, reasoning being after all?).

D. CONSCIOUSNESS, INFORMATION, AND PIZZA

"If one had to describe the deepest motivation for materialism, one might say that it is simply a terror of consciousness" (Searle, 1992, p. 55). Oh dear. If I had my way, I would give in to the terror and just not mention consciousness at all. But it is worth a word or two now (and see Appendix II) for two reasons. One is because it is all too easy to see the facts about conscious experience (the "second aspect of the problem of mindfulness" described in the Introduction) as constituting a knock-down refutation of the strongest version of the computationalist hypothesis. The other is because consideration of these issues helps to highlight important differences between informational and "merely physical" phenomena. So here goes.

How could a device made of silicon be conscious? How could it feel pain, joy, fear, pleasure, and foreboding? It certainly seems unlikely that such exotic capacities should flourish in such an unusual (silicon) setting. But a moment's reflection should convince you that it is equally amazing that such capacities should show up in, of all things, meat (for a sustained reflection on this theme, see the skit in Section 1.3). It is true, of course, that the only known cases of conscious awareness on this planet *are* cases of consciousness in carbon-based organic life forms. But this fact is rendered somewhat less impressive once we realize that all earthly life forms share a common chemical ancestry and lines of descent. In any case, the question, at least as far as the central thesis of the present chapter is concerned, is not whether our local carbon-based organic structure is crucial to all possible versions of conscious awareness (though it sounds anthropocentric in the extreme to believe that it is), but whether meeting a certain abstract computational specification is enough to *guarantee* such conscious awareness. Thus even the philosopher John Searle, who is famous for his attacks on the equation of mindware with software, allows that "consciousness might have been evolved in systems that are not carbon-based, but use some other sort of chemistry altogether" (Searle, 1992, p. 91). What is at issue, it is worth repeating, is not whether other kinds of stuff and substance might support conscious awareness but whether the fact that a system exhibits a certain computational profile is enough (is "sufficient") to ensure that it has thoughts, feelings, and conscious experiences. For it is crucial to the strongest version of the computationalist hypothesis that where our mental life is concerned, *the stuff doesn't matter*. That is to say, mental states depend solely on the program-level, computational profile of the system. If conscious awareness were to turn out to depend much more closely than this on the nature of the actual physical stuff out of which the system is built, then this global thesis would be either false or (depending on the details) severely compromised.

Matters are complicated by the fact that the term "conscious awareness" is something of a weasel word, covering a variety of different phenomena. Some use it to mean the high-level capacity to reflect on the contents of one's own thoughts. Others have no more in mind that the distinction between being awake and being asleep! But the relevant sense for the present discussion (see Block, 1997; Chalmers, 1996) is the one in which to be conscious is to be a subject of experience—to feel the toothache, to taste the bananas, to smell the croissant, and so on. To experience some *x* is thus to do more than just register, recognize, or respond to *x*. Electronic detectors can register the presence of semtex and other plastic explosives. But, I hope, they have no experiences of so doing. A sniffer dog, however, may be a different kettle of fish. Perhaps the dog, like us, is a subject of experience; a haven of what philosophers call "qualia"—the qualitative sensations that make life rich, interesting, or intolerable. Some theorists (notably John Searle) believe that computational accounts fall down at precisely this point, and that as far as we can tell it is the implementation, not the program, that explains the presence of such qualitative awareness. Searle's direct attack on computationalism is treated in the next chapter. For now, let us just look at two popular, but flawed, reasons for endorsing such a skeptical conclusion.

The first is the observation that "simulation is not the same as instantation." A rainstorm, simulated in a computational medium, does not make anything actually wet. Likewise, it may seem obvious that a simulation, in a computational

-based organic life
we realize that all
of descent. In any
present chapter is con-
-ture is crucial to all
anthropocentric in the
abstract computational
-ness. Thus even the
the equation of mind-
been evolved in systems
altogether" (Searle,
whether other kinds of
whether the fact that
is "sufficient") to en-
For it is crucial to
where our mental life
states depend solely on
conscious awareness were
nature of the actual
global thesis would be ei-
-sed.
"conscious awareness" is
phenomena. Some use
of one's own thoughts.
being awake and being
Block, 1997; Chalmers,
of experience—to feel
and so on. To experi-
or respond to *x*. Elec-
and other plastic explosives.
dog, however, may be
of experience; a haven
-ions that make life rich.
Searle) believe that com-
and that as far as we can tell
presence of such qual-
-ism is treated in the next
flawed, reasons for endors-

same as instantation."
does not make anything ac-
-tion, in a computational

medium, of the brain states involved in a bout of black depression will not add one single iota (thank heaven) to the sum of real sadness in the world.

The second worry (related to, but not identical to the first) is that many feelings and emotions look to have a clear chemical or hormonal basis and hence (hence?) may be resistant to reproduction in any merely electronic medium. Sure, a silicon-based agent can play chess and stack crates, but can it get drunk, get an adrenaline high, experience the effects of ecstasy and acid, and so on?

The (genuine) intuitive appeal of these considerations notwithstanding, they by no means constitute the knock-down arguments they may at first appear. For everything here depends on what *kind* of phenomenon consciousness turns out to be. Thus suppose the skeptic argues as follows: "even if you get the overall inner computational profile just right, and the system behaves just like you and I, it will still be lacking the inner baths of chemicals, hormones, and neurotransmitters, etc. that flood our brains and bodies. Maybe without these all is darkness within—it just looks like the "agent" has feelings, emotions, etc., but really it is just [what Haugeland (1981a) terms] a "hollow shell." This possibility is vividly expressed in John Searle's example of the person who, hoping to cure a degenerative brain disease, allows parts of her brain to be gradually replaced by silicon chips. The chips preserve the input–output functions of the real brain components. One logical possibility here, Searle suggests, is that "as the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior" (Searle, 1992, p. 66). In this scenario (which is merely one of several that Searle considers), your actions and words continue to be generated as usual. Your loved ones are glad that the operation is a success! But from the inside, you experience a growing darkness until, one day, nothing is left. There is no consciousness there. You are a zombie.

The imaginary case is problematic, to say the least. It is not even clear that we here confront a genuine logical possibility. [For detailed discussion see Chalmers (1996) and Dennett (1991a)—just look up zombies in the indexes!] Certainly the alternative scenario in which you *continue* your conscious mental life with no ill effects from the silicon surgery strikes many cognitive scientists (myself included) as the more plausible outcome. But the "shrinking consciousness" nightmare does help to focus our attention on the right question. The question is, just *what is* the role of all the hormones, chemicals, and organic matter that build normal human brains? There are two very different possibilities here and, so far, no one knows which is correct. One is that the chemicals, etc. affect our conscious experiences *only by affecting* the way information flows and is processed in the brain. If that were the case, the same kinds of modulation may be achieved in other media by other means. Simplistically, if some chemical's effect is, e.g., to speed up the processing in some areas, slow it down in others, and allow more information leakage between adjacent sites, then perhaps the same effect may be achieved in a purely electronic medium, by some series of modulations and modifications of current flow. Mind-altering "drugs," for silicon-based thinkers, may thus take the form of black-market software packages—packages that temporary induce a new pattern of flow and functionality in the old hardware.

There remains, however, a second possibility: perhaps the experienced nature of our mental life is not (or is not just) a function of the flow of information. Perhaps it is to some degree a direct effect of some still-to-be-discovered physical cause or even a kind of basic property of some types of matter (for extended discussion of these and other possibilities, see Chalmers, 1996). If this were true, then getting the information-processing profile exactly right would still fail to guarantee the presence of conscious experience.

The frog at the bottom of the beer glass is thus revealed. The bedrock, unsolved problem is whether conscious awareness is an *informational* phenomenon. Consider the difference. A lunch order is certainly an informational phenomenon. You can phone it, fax it, E-mail it—whatever the medium, it is the same lunch order. But no one ever faxes you your lunch. There is, of course, the infamous Internet Pizza Server. You specify size, consistency, and toppings and await the on-screen arrival of the feast. But as James Gleick recently commented, "By the time a heavily engineered software engine delivers the final product, you begin to suspect that they've actually forgotten the difference between a pizza and *a picture* of a pizza" (Gleick, 1995, p. 44). This, indeed, is Searle's accusation in a nutshell. Searle believes that the conscious mind, like pizza, just *ain't an informational phenomenon.* The stuff, like the topping, really counts. This could be the case, notice, even if many of the *other* central characteristics of mindware reward an understanding that is indeed more informational than physical. Fodor's focus on reason-guided state-transitions, for example, is especially well designed to focus attention away from qualitative experience and onto capacities (such as deciding to stay indoors when it is raining) that can be visibly guaranteed once a suitable formal, functional profile is fixed.

We are now eyeball to eyeball with the frog. To the extent that mind is an informational phenomenon, we may be confident that a good enough computational simulation will yield an actual instance of mindfulness. A good simulation of a calculator is an instance of a calculator. It adds, subtracts, does all the things we expect a calculator to do. Maybe it even follows the same hidden procedures as the original calculator, in which case we have what Pylyshyn (1986) terms "strong equivalence"—equivalence at the level of an underlying program. If a phenomenon is informational, strong equivalence is surely sufficient[4] to guarantee that we confront not just a model (simulation) of something, but a new exemplar (in-

[4]Sufficient, but probably not necessary. *x* is sufficient for *y* if when *x* obtains, *y* always follows. Being a banana is thus a sufficient condition for being a fruit. *x* is necessary for *y* if, should *x* fail to obtain, *y* cannot be the case. Being a banana is thus not a necessary condition for being a fruit—being an apple will do just as well.

stantiation) of that very thing. For noninformational phenomena, such as "being a pizza," the rules are different, and the flesh comes into its own. Is consciousness like calculation, or is it more like pizza? The jury is still out.

1.3 A Diversion

[This is extracted from a story by Terry Bisson called "Alien/Nation" first published in *Omni* (1991). Reproduced by kind permission of the author.]

"They're made out of meat."

"Meat?"

"Meat. They're made out of meat."

"Meat?"

"There's no doubt about it. We picked several from different parts of the planet, took them aboard our recon vessels, probed them all the way through. They're completely meat."

"That's impossible. What about the radio signals? The messages to the stars."

"They use the radio waves to talk, but the signals don't come from them. The signals come from machines."

"So who made the machines? That's who we want to contact."

"They made the machines. That's what I'm trying to tell you. Meat made the machines."

"That's ridiculous. How can meat make a machine? You're asking me to believe in sentient meat."

"I'm not asking you, I'm telling you. These creatures are the only sentient race in the sector and they're made out of meat."

"Maybe they're like the Orfolei. You know, a carbon-based intelligence that goes through a meat stage."

"Nope. They're born meat and they die meat. We studied them for several of their life spans, which didn't take too long. Do you have any idea of the life span of meat?"

"Spare me. Okay, maybe they're only part meat. You know, like the Weddilei. A meat head with an electron plasma brain inside."

"Nope. We thought of that, since they do have meat heads like the Weddilei. But I told you, we probed them. They're meat all the way through."

"No brain?"

"Oh, there is a brain all right. It's just that the brain is made out of meat!"

"So . . . what does the thinking?"

"You're not understanding, are you? The brain does the thinking. The meat."

"Thinking meat! You're asking me to believe in thinking meat!"

"Yes, thinking meat! Conscious meat! Loving meat. Dreaming meat. The meat is the whole deal! Are you getting the picture?"

"Omigod. You're serious then. They're made out of meat."

"Finally, Yes. They are indeed made out of meat. And they've been trying to get in touch with us for almost a hundred of their years."

"So what does the meat have in mind?"

"First it wants to talk to us. Then I imagine it wants to explore the universe, contact other sentients, swap ideas and information. The usual."

"We're supposed to talk to meat?"

"That's the idea. That's the message they're sending out by radio. Hello. Anyone out there? Anyone home? That sort of thing."

"They actually do talk, then. They use words, ideas, concepts?"

"Oh, yes. Except they do it with meat."

"I thought you just told me they used radio."

"They do, but what do you think is on the radio? Meat sounds. You know how when you slap or flap meat it makes a noise? They talk by flapping their meat at each other. They can even sing by squirting air through their meat."

"Omigod. Singing meat. This is altogether too much. So what do you advise?"

"Officially or unofficially?"

"Both."

"Officially, we are required to contact, welcome, and log in any and all sentient races or multi beings in the quadrant, without prejudice, fear, or favor. Unofficially, I advise that we erase the records and forget the whole thing."

"I was hoping you would say that."

"It seems harsh, but there is a limit. Do we really want to make contact with meat?"

"I agree one hundred percent. What's there to say?" 'Hello, meat. How's it going?' But will this work? How many planets are we dealing with here?"

"Just one. They can travel to other planets in special meat containers, but they can't live on them. And being meat, they only travel through C space. Which limits them to the speed of light and makes the possibility of their ever making contact pretty slim. Infinitesimal, in fact." "So we just pretend there's no one home in the universe."

"That's it."

"Cruel. But you said it yourself, who wants to meet meat? And the ones who have been aboard our vessels, the ones you have probed? You're sure they won't remember?"

"They'll be considered crackpots if they do. We went into their heads and smoothed out their meat so that we're just a dream to them."

"A dream to meat! How strangely appropriate, that we should be meat's dream."

"And we can mark this sector unoccupied."

"Good. Agreed, officially and unofficially. Case closed. Any others? Anyone interesting on that side of the galaxy?"

"Yes, a rather shy but sweet hydrogen core cluster intelligence in a class nine star in G445 zone. Was in contact two galactic rotations ago, wants to be friendly again."

"They always come around."

"And why not? Imagine how unbearably, how unutterably cold the universe would be if one were all alone."

1.4 Suggested Readings

For an up-to-date, and indeed somewhat sympathetic, account of the *varieties of dualism*, see D. Chalmers, *The Conscious Mind* (New York: Oxford University Press, 1996, Chapter 4).

For *general philosophical background* (identity theory, behaviorism, machine functionalism) a good place to start is Appendix I of this text and then P. M. Churchland, *Matter & Consciousness* (Cambridge, MA: MIT Press, 1984, and subsequent expanded editions). Another excellent resource is D. Braddon-Mitchell and F. Jackson, *Philosophy of Mind and Cognition* (Oxford, England: Blackwell, 1996, Chapters 1, 2, 3, 5, 6, and 7).

For the *broad notion of a computational view of mind*, try the Introductions to J. Haugeland, *Mind Design*, 1st ed. (Cambridge, MA: MIT Press, 1981) and *Mind Design II* (Cambridge, MA: MIT Press, 1997). The former ("Semantic engines: An introduction to mind design") is especially good on the syntax/semantics distinction, and the latter ("What is mind design?") adds useful discussion of recent developments.

For more on *Turing machines*, see J. Kim, "Mind as a computer," [Chapter 4 of his excellent book, *Philosophy of Mind* (Boulder, CO: Westview Press, 1996)]. Chapters 1–3 cover *dualism, behaviorism, and identity theory* and are also highly recommended. Chapter 4 focuses on the advent of *machine functionalism* and includes detailed discussion of the antireductionist themes that surface as the "structure not stuff" claim discussed in our text.

For *philosophical accounts of machine functionalism, and critiques*, see H. Putnam, "The nature of mental states." In H. Putnam (ed.), *Mind, Language & Reality: Philosophical Papers*, Vol. 2 (Cambridge, England: Cambridge University Press, 1975) (a classic and very readable account of machine functionalism) and N. Block, "Introduction: What is functionalism?" and "Troubles with functionalism." Both in his *Readings in Philosophy of Psychology*, Vol. 1 (Cambridge, MA: Harvard University Press, 1980). (Clean and critical expositions that nicely reflect the flavor of the original debates.)

J. Searle, "The critique of cognitive reason," Chapter 9 of his book, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992) is a characteristically direct *critique of the basic computationalistic claims* and assumptions.

A useful, *up-to-date introduction to the empirical issues* is S. Franklin, *Artificial Minds* (Cambridge, MA: MIT Press, 1995), and an excellent general *collection of papers* may be found in J. Haugeland, *Mind Design II* (Cambridge, MA: MIT Press, 1997).