1

What's in Your Mind?

Zenon W. Pylyshyn

1 Introduction

Neuropsychologists have an advantage on us dry cognitive scientists: They always have impressive color slides of PET or MRI or fMRI images showing the exact location of whatever they wish to discuss – the soul or the locus of sinful thoughts or the center of consciousness. If one were to go by popular science articles on the brain, one would have to conclude that we know where everything is located in the brain, and therefore we know everything about it except *how* it manages to do things like think. But I chose the title deliberately, because I believe that what we do here at the Center for Cognitive Science is precisely that we study what is in the mind. Let me explain.

The term "mind" has been associated with psychology at least as far back as William James, who defined psychology as the "Science of Mental Life." Yet in the past 50 years it has fallen into disfavor. But there is good reason to believe that this was a mistake and that psychology really is about the mind, and in particular that explanations of behavior must take into account what is in the mind. The question of what's in the mind should be answered in psychology in the same way that the parallel question is answered in physics. There, a question such as what's in this table or what's in the sun is answered by looking for properties, entities, and causal laws which explain the important regularities that define that particular science.

This essay is based in part on a lecture in the "What is Cognitive Science?" series held at the Rutgers Center for Cognitive Science in the fall of 1996.

The trouble with saying that answers to psychological questions should be provided by looking for properties and laws that explain important regularities in the science is that we do not know in advance of the development of the science exactly what will count as relevant regularities. This is a point that often escapes the social sciences. Physics does not consider it a failure if it can't explain why some parts of the table are dustier than others, why some parts are rougher, why some parts are warped, or why the wood will eventually rot because of bacteria and other microorganisms that infest it. It simply turns out that those are not the regularities that physics is equipped to answer. It doesn't even have categories like corner or rough or smooth or rotting in its laws. That's why there are other sciences, like microbiology perhaps, which address regularities based on such categories.

Notice that we do not demand that the terms that occur in the answer to the question "What's in this table?" be ones that we have any prior understanding of or expectations about, or even that they be things that we can see or feel or taste or otherwise have any sensory contact with. In psychology we always feel that we can set two kinds of agendas in advance. One is that we can say what the relevant data will be. For example, we say that psychology is in the business of predicting behavior. If this were true, much of our work would already be done, since there is already a way of predicting such behavior as that when people fall off the top of a building they accelerate at roughly 10 meters per second for every second of flight. But that's not psychology, you say! Exactly! But what exactly does count as psychology? The second, and closely related agenda that we often feel we can set in advance is specifying what the vocabulary or the categories will be in the science as it develops – as well as what sorts of things it will be able to explain. Is it to be concerned primarily with, say, voluntary behavior? That already presupposes that the category "voluntary" will be recognized and will play a role in the science of mind. Also "voluntary" is far from being a neutral term, since it assumes that we know what it is for some behavior to be voluntary. Moreover, it assumes that this is the type of behavior that cognitive science will be concerned to explain. It turns out that categories such as "voluntary" and "conscious" are very likely ones that we may have to give up as the science of mind develops a scientific base. Similarly, it has been widely assumed that psychology should be concerned with explaining "learning." But can we stipulate that in advance? Do we know what kinds of changes in behavior constitute learning, in the sense relevant to psychology (e.g., is the growth of hair and fingernails a type of "learning" and if not, why not?) and whether these changes will fall to psychology or biology or some other science to explain?

1.1 What is special about intelligent behavior?

The most remarkable property of human behavior involving intelligence (as well as similar behavior of certain other species), is that, in order to capture what is systematic about it, it is necessary to recognize equivalence classes of causal events that cannot be characterized using the terms of existing natural sciences. The anthropologist Kenneth Pike once made the astute observation that human behavior cannot be understood in terms of objective physical properties of the world, which he called *etic* properties, but only in terms of the way in which the world is perceived or represented in the mind, which he called *emic*, or internalized, properties. When viewed in terms of objectively defined classes of stimuli and responses, human behavior appears to be essentially stimulus-independent, and the attempt to cast it in terms of objectively defined stimulus properties runs up against either obvious counter-examples or self-contradictions (see, e.g., Chomsky's review of Skinner's attempt to do just that in his behaviorist analysis of language). On the other hand, when cast in terms of such constructs as beliefs and desires, and when reasoning is allowed as part of the process intervening between stimuli, representations, and actions, the picture becomes much more coherent (though still highly incomplete).

Consider typical folk-psychology explanations of ordinary behavior. Such explanations say, for example, that people do things because of what they know or believe and because of what they want, or more precisely because of their goals and utilities. Although such a general claim should be obvious, it has in fact been widely denied throughout the history of the field. The trouble with denying this truism is that without it you cannot explain the simplest piece of behavior, such as, for example, why there are people in the audience here today. You and I and your granny know that the reason there are people here is that they have been led to believe that there would be a talk given at this particular time in this room. Moreover, this is not a rough and approximate way of talking; it's really and truly the case. The way you know that it is truly the case is to consider what would have happened if the antecedent conditions in my explanation for why you are here had not been true – i.e., if you did not have the beliefs I said you had. For example, if you did not know that the talk was to be in this room, or did not know the time of the talk, or if you had some reason to discount the announcement that there would be a talk given here – for example, if you found out that I had not arrived in time, or if you had been led to believe through any of an indefinite number of ways, that the announcement you received was in error or that it was all a practical joke or that the building had to be evacuated because of a bomb scare, and so on and on without limit – and if I had reason to believe that *you* would not be here, then I too would not be here.

How often do you get such reliable predictions in scientific psychology? Notice that you only get such predictions if the explanatory vocabulary contains at least some of the terms of folk psychology – at least terms like "believes," along with terms for the contents of beliefs like "meeting," "talk," or even "practical joke" or "bomb scare." Moreover, you only get the predictions to come out if the beliefs and the meanings of sentences that people hear can enter into a certain kind of process, a process which we generically refer to as *inference*, wherein new beliefs are established that were not part of the original stimulus information, or, to put it differently, consequences are somehow drawn from the initial beliefs, goals, and data provided to the individual. It's absolutely clear that you cannot get by in cognitive psychology without, at the very minimum, having some way of dealing with this basic fact. Not even the most radical behaviorist fails to accept this fact, even though the polemics very often deny it.

So, for instance, while denying that meanings and knowledge and goals are relevant to the prediction of behavior, behaviorists still make use of the fact that they can predict people's behavior by putting up a poster containing sentences whose meaning is, for example, that if a person shows up at a certain time to take part in an experiment, that person will be paid a certain sum of money or will receive credit toward a course requirement. Notice that the experimenter implicitly accepts that the form of words on the poster, or its physical layout, is not what is relevant to predicting the reader's behavior; what matters is that the poster contains sentences with a certain meaning for the intended readership, and that in the proper context, rational people would come to have certain beliefs after reading those sentences, and that those beliefs together with the readers' goals and utilities would lead them to act in a certain way.[1]

1.2 Meaning and causality

The point of the subject-soliciting poster example is this: The relevant equivalence class of stimuli needed to predict behavior is the class of *synonymous sentences*, or the class of sentences that *mean* the same thing or at least underwrite the same belief. But this equivalence class contains an unbounded number of stimuli, and what the members of this class have in common cannot be specified physically – being "synonymous" is what is called a semantic property. What distinguishes one science from another is the class of events or properties that they appeal to. Geology talks about mountains and rivers; economics talks about value and supply and demand; meteorology talks about precipitation and storms and the jet stream, and so on. In each case the things being discussed are physical things, but the categories are not the ones that physics recognizes – and they differ from science to science. Psychology needs to speak of how we perceive a stimulus, what we believe and what we want – or, more generally, how we *represent* the world (see below).

Now if you accept this – and it would be irrational not to – then you are led immediately to ask how it is possible for a biological entity made of protoplasm and governed by natural laws to have such a property. And that's where the trouble begins, for this is a highly non-trivial problem for a number of reasons. Here is a glimpse of one such reason. In every science, when you have an explanation of the form "Y occurs because of X," then anything that fills the slot X is a causal property; hence any property mentioned in that statement must on each occasion have a real physical existence (assuming, as most of us do, that only physical things can serve as causes). But what about the explanation that you came to this room at this time because you believed there would be a talk given here? It is intrinsic to the explanation that it mention a talk. Yet the explanation would continue to be true whether or not there was in fact a talk. All that is required is that you *believed* that. But that makes belief a strange sort of property, a property characterized in terms of something that need not exist!

It is a true explanation of why King Arthur's knights did certain things in the Middle Ages that they were in search of the Holy Grail, or that other people did certain things because they were searching for the pot of gold at the end of the rainbow. And those explanations hold whether or not there is such a thing as a Holy Grail or a pot of gold at the end of the rainbow. The same is true of people who do things in order to win someone's love or to gain tenure. Depending on what the beliefs are about, people act differently, though nonetheless appropriately to the content of their beliefs. Beliefs about different things count as different beliefs. And this is true whether or not what the beliefs are about exists in the world, or whether it is even physically possible for it to exist (e.g., ghosts). How, then, can the content of belief enter into the causation or the explanation of behavior?

Needless to say, this is a venerable old puzzle, one that was first brought to the attention of psychologists by Franz Brentano, and one which is still hotly debated by philosophers. But it is fair to say that within the research community that identifies with cognitive science and artificial intelligence, there is a hypothesis that has become so deeply entrenched that it is simply taken for granted. The hypothesis is this. What makes it possible for systems – computers or intelligent organisms – to behave in a way that is correctly characterized in terms of what they represent (say, beliefs and goals) is that the representations are *encoded* in a system of physically instantiated symbolic codes. And it is because of the physical form that these codes take on each occasion that the system behaves the way it does, through the unfolding of natural laws over the physical codes.

Stated in this bald way, this may sound like an esoteric and philosophical doctrine. But there is one thing that makes this story more than a little plausible, and that's the fact that it is clearly and literally true of computers. It explains why a computer can be correctly described as behaving in a certain way because of what it represents (e.g., it contains knowledge about medical symptoms and their etiology and is told what symptoms a person has, so it infers a diagnosis and suggests medications). Without getting into the more controversial aspects of the claim that this is the correct way to describe what the computer is doing, it is at least an existence proof that it is possible to have a system which is both clearly governed by physical laws, and at the same time whose behavior can be given a coherent account in terms of what it represents.

1.3 Symbols, codes, and computing

There is good reason why computers can be described as processing knowledge. The reason was discovered at around the same time as the idea of computing itself was developed. This discovery came, perhaps surprisingly, from the development of mathematics and logic in the first half of the twentieth century. A number of far-reaching mathematical ideas came together in the 1930s, associated with names like Hilbert, Kurt Gödel, Betrand Russell (with Alfred North Whitehead), Alan Turing, Alonzo Church, and other logicians. The discovery was this: Reasoning about meaningful things – about things in the world or in the imagination – could be carried out by a process that itself knew nothing of the world or of meanings, did not know what its "thoughts" were about!

To illustrate this fundamental idea, consider what is involved when you go from a set of beliefs to a new belief. Suppose you know (somehow) that John is married either to Mary or to Susan. Then suppose you discover that John is in fact not married to Susan. You can then conclude that he must be married to Mary. We can represent this by equations such as the following, which involve (in this case) two special terms, called Logical Terms, "*or*" and "*not.*"

(1) Married (John, Mary) *or* Married (John, Susan)

and the equation or "statement"

(2) *not* (Married (John, Susan))

From these two statements you can conclude,

(3) Married (John, Mary)

But notice that (3) follows from (1) and (2) *regardless* of what is in the parts of the equation not occupied by the terms *or* or *not* so that you could write down the equations without mentioning marriage or John or Mary or, for that matter, anything having to do with the world. Try replacing these expressions with the meaningless letters ***P*** and ***Q***. The inference still holds:

(1′) ***P*** *or* ***Q***
(2′) *not* ***Q***

Therefore,
(3′) ***P***

The idea that logical inference can be carried out by a process of examining meaningless symbols leads directly to the foundational assumption of cognitive science, which is that thought is a species of computing. This is because the sort of "meaningless" manipulation of symbols just described is just what computers are good at. So if the idea is correct, maybe computing is what the brain does to produce intelligent behavior. The bridge from formal symbol manipulation to computing was completed in 1936 by the mathematician Kurt Gödel who showed that anything that could be described in terms of manipulations of symbols could be carried out by a very simple machine (later called a Turing machine), which became the defining property of reasoning and later of intelligent action.

2 The Tri-level Hypothesis

The behavior of complex systems can often be described at different levels. Sometimes this may be just a convenience in talking about them (e.g., we can describe a car at various levels of specificity). But sometimes this is essential, because the system really has different levels of organization. For example, there appears to be a level of organization at which the laws of economics, like Gresham's law or the law of supply and demand, hold. These are genuine, principled levels at which certain organizing principles apply. If we could describe only the movement of currency and goods, we would have no hope of discovering principles of economics, because the principles hold regardless of what physical form "money" and "goods" take. We all know now that transfers of funds can take place by the most exotic means, including codes sent over a digital network, and that goods and services can also take the most surprising forms; yet the principles of economics and the laws of contractual obligation hold irrespective of the forms that goods, services, payments, and contractual transactions take.

When it comes to trying to understand cognition, the current view in cognitive science is that there are at least three distinct levels at which intelligent systems are organized (this is the so-called tri-level hypothesis discussed at length in my 1984 book – see note 3):

1 The biological or physical level
2 The symbolic or syntactic level
3 The knowledge or semantic level

What this proposal amounts to is the claim that there are different generalizations that exist at each of these levels. There are patterns of behavior that

Figure 1.1 Electromechanical calculator. How do you explain different aspects of its behavior?

can only be explained by appeal to biology – for example, why people's reactions slow down when they drink alcohol, why they get irritated when deprived of sleep, why their memories worsen with age, why certain behaviors change at puberty, and so on. We have already seen that some patterns of behavior can be explained only by appeal to what people want and what they believe (we will see in the next section that the semantic levels also takes in a wider range of behaviors than just rational decisions, since many of the organizing principles of perception, memory, and other aspects of cognition also require that we refer to how aspects of the world are represented – which makes intelligent behavior special in being "representation-governed"). The new twist in the tri-level picture is the idea that the knowledge level is implemented through a system of codes, more or less as we discussed in the previous section.

The idea that different phenomena may require that we appeal to principles at different levels is already familiar to us, since it is routine in computing. For example, a computer may implement an economic model. If it fails to make the correct prediction of a certain change in the economy, we explain that by reference to economic factors, not to properties of the computer program itself, or to the electronics of the computer. But there are cases when we might indeed explain the model's behavior by reference to the program itself – for example, if the program had a bug in it. Similarly there are situations (e.g., a power failure) such that we would explain the behavior by appealing to the electronics. The situation is also clear in the case of a calculator, such as the one shown in figure 1.1.

Various questions can be asked about the calculator's behavior:

1 Why is the calculator's printing faint and irregular? Why are parts of numbers missing in the LED display?
2 Why does it take longer to calculate large numbers than small ones?
3 Why does it take longer to calculate (and display) trigonometrical functions (such as sine and cosine) than sums?
4 Why does it take longer to calculate the logarithms of large numbers than

of small numbers, whereas it takes the same length of time to add large numbers as to add small numbers?

5 Why is it especially fast at calculating the logarithm of 1?

6 Why is it that when one of the keys (labeled √) is pressed after a number is entered, the calculator prints what appears to be the square root of that number? How does it know what the square root is?

7 How does the calculator know the logarithm of the number I punch in?

8 When the answer to an arithmetic problem is too long to fit in the display window, why does the form of the answer change and some of the digits get left off?

9 Why is it that even when the answer fits in the window, some of the right-hand digits in the answer are different from what I get when doing it by hand? (It is sometimes off by 1).

It is clear that different *kinds* of answers apply to these questions.[2] Some require an answer stated in terms that refer to electrical and mechanical things – they require physical-level explanations (e.g., question 1). Others require symbol-level explanations – for example, they require that one describe the "method" or algorithm used by the calculator (e.g., questions 2–7 require such a description), including (for question 5) whether some of the answers are pre-computed and stored. Others require something in between the symbol level and the physical level; they require that we describe the machine's *architecture* – which is to say, we must describe things such as the *size of the storage registers* it uses. Notice that the size (in terms of number of bits, or bytes) is not a physical property, since the answer would apply to calculators that were physically quite different (you could ask about the register size of your PC, which works quite differently from the calculator). Questions 8 and 9 concern what are called "rounding errors," and the answer would need to address how individual numbers are represented and what principle applies when the capacity of a register is exceeded. The principle may well reside in the design of the architecture of the calculator, and not in the program it uses in a particular case.

Several of the questions are actually about the relation between the calculator and the world of abstract mathematics. Saying that the calculator computes the *sine* or *logarithm* function is to say more than just what algorithm it uses. It is to claim that the algorithm in question actually computes representations of numbers that correspond to certain mathematically defined abstract functions. Showing that this is the case can be a difficult task. Mathematicians are sometimes concerned to prove mathematically that a certain program will *always* generate outputs that are consistent with a certain mathematically defined function (even though it can only be tested on some finite subset of inputs). This is the computer science task of proving the correctness of programs – a difficult and challenging problem in theoretical computer science. In order to do this, the theorist needs to describe the computer's operation in terms of the mathematical function it was designed to compute.

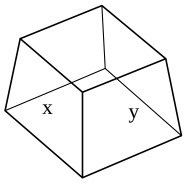

Figure 1.2 Reversing wire figure, showing "coupling" of perceived properties.

In other words, for purposes of proving correctness, the machine must be described in terms of the things it represents (abstract mathematical objects) – this is the semantic level of description.

2.1 Representation-governed behavior

The idea, sketched above, that certain behavioral regularities can be attributed to different representations (some of which are called "beliefs" because they enter into rational inferences) and to symbol-manipulating processes operating over these representations, is a fundamental assumption of cognitive science. This idea is an instance of what is a fundamental claim about intelligent systems: Intelligent systems (including animals and computers) are governed by *representations.* To explain the simplest fact about the behavior of most "higher" organisms, we must say how some aspect of the world is represented – and this applies even where the behavior does not appear to involve reasoning or rational decision making. For example, it is a remarkable fact about perception that you can only state the generalizations or laws of perceptual organization in terms of how patterns are perceived, not in terms of their physical properties. Here are some examples that should make this clear.

Consider the "laws" of color mixing. When yellow light is mixed with red light, the resulting light appears orange. Is this a law about how different wavelengths are perceived? The answer is no. There is an unlimited variety of ways of producing yellow light (by filtering white light to allow only wavelengths of 580 nm or by mixing light of other wavelengths such as 530 nm and 650 nm). Similarly, there is an unlimited variety of ways of producing red light. But *regardless of how each light is produced*, mixing the two lights produces a light that *looks* orange – providing only that one of the lights *looks* yellow and the other *looks* red! How some aspect of a percept looks depends not on objective properties of the display, but on how parts of the display appear. Another way to say this is that how something is seen depends on how different aspects of it are seen or are *represented* by the perceiver. In figure 1.2 above, how the object is seen depends on how you see its parts. If you see edge X as part of the nearest face, then you will also see edge Y as part of the nearest face and the vertex where these two meet as the upper corner nearest you. In that case you are also likely to see the face bounded by X and Y as

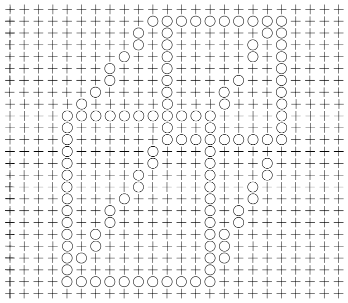

Figure 1.3 An alternative way to present the cube figure.

being the same size as the other faces – that is, you are likely to see the figure as a cube. But if you see the face formed by X and Y as the bottom of a figure (seen from above) then that face is likely to be seen as larger than the top face – so the figure looks like a cut-off pyramid.

This sort of "coupling" between how parts of a figure are perceived is an extremely general phenomenon. The organizing principle cannot be stated over the geometrical properties of the figure, only over its *perceived* properties – or, in our terms, over how parts are *represented* by the mind. The principles take the form, "If X is *represented as* (e.g., being closer, or being yellow, or . . .), then Y will be *represented as* . . . ." It is important to realize that such principles apply *no matter how the parts came to be represented the way they are* – exactly as was the case in the color-mixing example. There is an unlimited variety of ways of producing the perception of a certain edge or vertex. For example, it can be produced by drawing a line, or by selecting a set of pixel points from among an array and making them distinct, as in figure 1.3 where the subset of points defining the figure are distinct because of the shapes of the elements. And the subset can also be made distinct by jiggling the relevant elements while the other elements remain stationary, or by showing the array in stereo with the subset in a different depth plane, or by moving a narrow slit back and forth over the figure 1.3 so that only a slit is visible at any instant but the figure is still seen as lying behind the screen with the slit, and so on. Once again, it matters not how the information is presented. What matters is how it is *seen* or *represented.* This feature of intelligent processing, wherein *what is relevant to principles of operation is how something is represented*, is the main reason why we believe that intelligent processing is computational. In computation, it is how something is encoded – its symbolic form – that determines how the process runs. The computer does not "know" anything about the outside world: All it knows are the symbolic codes or data structures with which it deals.

2.2 What kind of computer is the mind?

If you buy the story I have been sketching, then you are ready to accept the general conclusion that the mind is a type of computer – a story that is getting to be easier and easier to accept in this day and age, when "Artificial Intelligence" is much discussed. But that still leaves a gaping hole in the cognitive science project: To specify what *kind* of computer the mind is. Notice that we are not claiming that the mind runs like your PC or MAC. Whatever kind of computer it is, it is clearly not *that* kind of computer. There is reason to believe that it does not always execute one instruction after another, that it does not store information by encoding it in terms of a binary pattern of bits, that it retrieves it by specifying the address where it is stored, and so on. What it does share with PCs and with all known forms of general-purpose computers is that it manipulates symbolic codes. There is much that needs to be said about even this claim, but such a discussion is beyond the scope of this essay.[3] But the project of understanding the nature of mind cannot get off the ground unless we take seriously the task of specifying, at least in general outline, what kind of computer the mind is. The reason why this is so central a task is itself revealing, and I will devote the rest of this chapter to spelling out the answer to the question, Why do we need to know what kind of computer the mind is?

The reason we need to understand what kind of computer the mind is, is that merely simulating intelligent functions – however interesting and difficult it may be – is not enough for the purpose of *explaining* human intelligence. That's because it is often possible to produce some piece of intelligent behavior in a manner totally different from how it is produced by the human mind. A good example of this is arithmetic. Computers can (and do, routinely) carry out arithmetic operations in a completely different way from the way you were taught in school – because it is faster and more convenient for them to use their special-purpose operations (e.g., using binary arithmetic, shift operations, and so on). The basic operations available to the computer (as well as the way it encodes, stores, and retrieves information) constitute what computer scientists call its functional or computational *architecture*. If we wanted to model how you and I do certain kinds of arithmetical problems, we would need to first find out what the computational architecture of the mind is (what we call its *cognitive architecture* – determining which is the most fundamental problem in all of cognitive science). The cognitive architecture is what determines what the mind can do and the way it can do it. It places strong constraints on any theory of cognitive functioning.

When we carry out some mental operation (say, solve some problem), we use two kinds of resources. One thing we use is what we know – what we have been told or what we have found out by reading or talking to people or by drawing inferences from what we already know to new knowledge. The second resource we use is our *cognitive capacity*: the capabilities that our brain

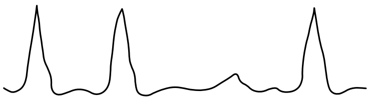

Figure 1.4 A typical pattern produced by an unknown box.

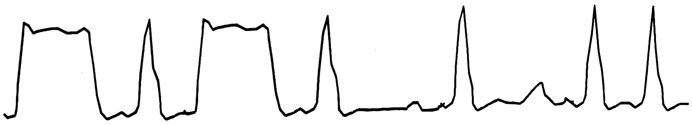

Figure 1.5 An exception to the typical pattern that ocurs in the special context show.

affords us. Much of this capacity is probably innate, but some of it may be acquired through maturation, practice, or other mechanisms we still do not understand. But this cognitive capacity is what we have because of our cognitive architecture, because of the kind of mind we have. The combination of what we know and what our capacities are is what determines what we do.

3 Cognitive Capacity

The idea of cognitive capacity, or cognitive architecture (I use the terms interchangeably), is a straightforward application of an idea from computer science. Because of this, it merits some examples to make it more concrete. What I will do is provide some very simple-minded examples to illustrate the following point: Merely predicting behavior is not good enough for purposes of explanation. We must also separate two major determinants of behavioral regularities: knowledge to highlight the distinction between a structurally defined capacity and a "mere" regularity) to one that is more relevant to a problem in cognitive science that has preoccupied me over the years. First the simple made-up example.

Suppose you were to find a mysterious box with unknown contents that was carrying out some function (also initially unknown) in its normal environment. Suppose further that the box had some conspicuous wires coming out of it that looked to be providing its normal behavioral "output." If we attach the wires to a recorder, we find that the box generates a variety of patterns of electrical activity in the course of its normal functioning. Among the patterns it generates are some single and some double spikes, as shown in figure 1.4.

As we examine the behavior of the box more carefully, we find that while the pattern of single and double spikes is usually as shown above, there are occasional exceptions in which the single spike precedes the double one. Such exceptions, however, occur in a predictable context. We find that the reverse pattern to that of figure 1.4 occurs only when it is preceded by two special long–short blip pairs, as in figure 1.5.

Let us assume that this pattern is quite reliable (we observe it over a long period of time). The question is: What does this pattern tell us about the nature of the box? Suppose you were to develop a theory of how the box works – say, by constructing a computer model that simulates its function. It would be very easy to do so, since the behavioral repertoire is quite simple. But what would we know about the nature of the box from such a model? Or, put another way, What does the behavioral regularity tell us about how the box works?

The answer is *nothing*. In this case, knowing the pattern of behavior tells us very close to nothing about how the box works. That's because we have observed only in its "typical" context or its "ecological niche," so cannot be aware that its capacity is far greater than is shown in that sample. I can reveal to you (because I made up the example!) that the box exhibits the pattern it does because of what the electrical patterns *represent*, not because of how the box is constructed. I can now tell you that the box is a device that transmits English words in International Morse Code (IMC). In IMC a single spike represents the letter *e*, a double spike represents the letter *i*, and the double long–short pattern represents the letter *c*. Thus the pattern we have observed arises entirely from a spelling rule in English: namely, "*i* before *e* except after *c*"! We can determine that the regularity in question does not arise from the architecture of the device even without knowing how it is constructed by simply observing that in different situations (not different wiring or a different physical arrangement) the behavior would be quite different. For example, if we got it to transmit words in German or French, the regularity would disappear. Observing this sort of change in behavior without changing the system's physical structure is one of the main methodological tools we have for distinguishing architectural from representational determinants of behavioral patterns. We will see more of this methodological tool below, where the informational alteration of behavioral regularities is called *cognitive penetration*.

The message of the above example (and other examples I will present below) is that when you encounter a systematic pattern of behavior (what I have called a "regularity" or a "generalization"), you need to ask *why* that generalization holds: Is it because of the way the mind is, or is it because of what we *know* or how we represent the world – because of the architecture or because of properties of what is represented.

Here is another example. Understanding natural language is one of humans' unique and most important and fluent skills. There have been many studies showing complex, sophisticated computations performed in the course of understanding a sentence. Some of the operations we perform on parts of a sentence (such as looking up a string of characters in a mental dictionary, or "lexicon," to check on what concept it corresponds to and what grammatical form it might take) may reveal properties of the cognitive architecture. But some don't. Take, for example, the pair of sentences below and ask yourself who the italicized pronouns refer to in each case. Then ask whether the answer reveals something about the architecture of the language-understanding system or whether it reveals something about what the listener knows about the world.

(1) John gave the book to Fred because *he* finished reading it.
(2) John gave the book to Fred because *he* wanted to read it.

In this case we would expect the explanation of why the pronoun refers to different people in the two sentences to appeal to one's knowledge of what books are for and where things end up when they are given. Only factors like this would explain why in particular cases the pronouns are assigned different referents in the two sentences and why the reference assignment could be easily changed in a logically coherent way by altering the belief context. (For example, suppose we knew that John was trying to encourage Fred to learn to read and had promised Fred the book as a reward if Fred finished reading all of it; or if we knew that John was blind and that Fred would often read to him. In such cases we might well assign the pronouns differently in these sentences.) In other words, the cognitive penetrability of the observed regularity marks it as being knowledge-dependent and as involving reasoning – even if one is not aware of such reasoning taking place. It is within the cognitive capacity of the organism to assign a different referent to the pronoun, with the new assignment being explicable in terms of the principles that explained the original assignment – namely, in terms of an inference from general background beliefs. The difference between the cases would be attributed to a difference in the state of knowledge or belief, not to a difference in their capacity or cognitive architecture.

Let's look at a somewhat different case that is of special interest to us as psychologists or students of cognitive functioning. It is known, through countless experiments, that when people imagine certain situations, they tend to exhibit many patterns of behavior (particularly of timing) that are similar to those that would be observed if they witnessed the corresponding situation. For example, it takes longer to "see" details in a "small" mental image than in a "large" image; it takes longer to imagine solving a construction problem (such as the task of folding a piece of paper to form a given figure) if it would have taken you a large number of steps to solve it in real life; and so on. As noted above, in order to decide whether this is due to the architecture or to the represented world, we need to ask why each of these regularities holds.

Take, for example, the case of "mental color mixing" at which many people excel. Suppose I ask you to imagine a transparent yellow disk and a transparent red disk, and then to imagine that the two disks are slowly moved together until they overlap (as in the color-mixing example mentioned earlier). What color do you see now where they overlap? People differ in their abilities to imagine color mixing. But no matter what color you see the overlapping disks to be, or whether you even experience any color mixing at all, the question of interest to us is *why*: Is the color you see in this example a result of the nature

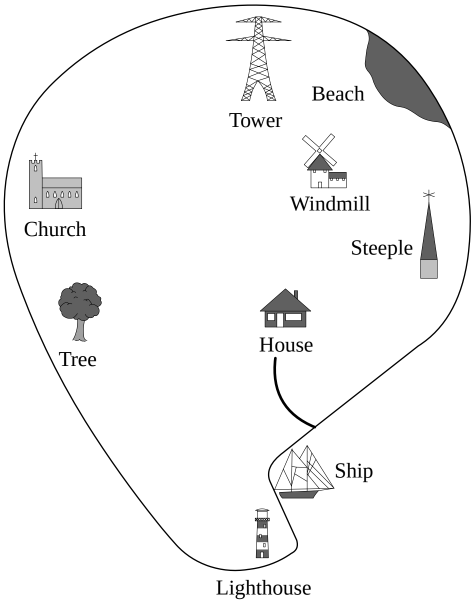

Figure 1.6 Map to be learned and imaged in one's "mind's eye" to study mental scanning.

or structure of your mind, or is it a result of what you know or remember about how colors mix? In this case it seems clear enough that what determines the solution you get is not some intrinsic property of your mind or brain – its cognitive architecture – but your memory or knowledge of how things work in the world and in perception. Of course, being able to remember such laws or past experiences *is* a property of your mind, but the actual way that colors mix (in contrast with what happens in vision, where the laws of color mixture really are immutable properties of the visual system) is very likely not, since you can make the overlapping pair of filters have any color you want it to have.

Here is another similar example of using a mental image – only this one represents a real experimental finding that has received a great deal of attention in the psychological literature on mental imagery. It has been shown over and over that it takes you longer to "see" a feature in an image if that feature is further away from one you have just examined. So, for example, if you are asked to imagine a dog and inspect its nose and then to "see" what its tail looks like, it will take you longer than if you were asked to first inspect its hind legs. Here is an actual experiment carried out by Steven Kosslyn.[4] Subjects were asked to memorize a map such as the one in figure 1.6.

They were then asked to imagine the map and to focus their attention on one place – say the "church." Then the experimenter said the name of a second place (say, "beach" or "ship") and subjects were asked to press a button as soon as they could "see" the second named place on their image of the map. What Kosslyn (and many others) found is that the further away the second place is from the place on which subjects initially focused, the longer it takes to "see" the second place in their "mind's eye." From this, most researchers have concluded that larger map distances are represented by greater distances, in some mental space. In other words the conclusion is that mental images have spatial properties – they have *magnitudes.* This is a strong conclusion about cognitive architecture. It says, in effect, that the symbolic code idea we have discussed earlier does not apply to mental images. In a symbolic encoding two places can be represented as being further away just the way we do it in language; by saying the places are *x* meters (or whatever) from one another. But the representation of larger distances is not itself in any sense *larger*. The question, then, is: Is this conclusion about architecture warranted? Does the difference in time in this case reveal a property of the architecture or a property of what is represented. This exactly parallels the situation in the code-box example, where we asked whether a particular regularity revealed a property of the architecture or a property of what was being represented. In that case the fact that the regularity shown in figures 1.4 and 1.5 goes away if the box transmits words in another language suggests that it is not. What about the image-scanning case? Is it like the code-box case or the imagined color-mixing case, or does the time pattern indicate something about the architecture, as generally assumed? To answer this question, we need to determine whether the pattern arises from a fixed capacity of the image-encoding system or whether it can be changed by changing the task or the beliefs people hold about how things work in the world.

This is a question to be settled by careful experiment. But there is already informal reason to suspect that the time course of scanning is not a property of the cognitive architecture. Do the following test on yourself. Imagine that there are lights at each of the places on the imagined map. Now imagine that a light goes on at, say, the beach. Now imagine that this light goes off and one comes on at the lighthouse. Did you need to scan your attention to see this happen and to see the lit-up lighthouse in your "mind's eye"? We did this experiment by showing subjects a real map with lights at the target locations. We allowed them to turn lights on and off. Whenever a light was turned on at one location, it was simultaneously extinguished at other locations. Then we asked subjects to imagine the map and to indicate (by pressing a button) when a light was on and they could see the illuminated place. The time between button presses was recorded and correlated to the distances between illuminated places on the map. As expected, the result was that there was no relation between distance on the imagined map and time. You might think: Of course there was no time increase with increasing distance, because I was not imagining that I was scanning that distance. That's just the point: You can imagine scanning the imagined map if you want to, or you can imagine just hopping from place to place on the imaginary map. If you imagine scanning, you can imagine scanning quickly or slowly at a constant speed or at some variable speed. You can in fact, do whatever you wish, since it is *your image* and *your imagining*, so you can make it do whatever you like over time! If the marchitecture restricts the operations you can perform (which it may well do), this does not show up in the experimental data on timing that are widely cited as showing that images are laid out in space. Thus it appears that the time pattern of mental scanning is like the pattern of blips observed in the code-box example. In both cases, while the pattern *could* have been due to the architecture of the relevant system, the evidence we discussed suggests that it is not. Rather, it is due to a correctly encoded pattern in the represented domain. In the code-box case this pattern is due to the spelling of English words, and in the image-scanning case it arises from the fact that subjects know what happens when you scan a picture with your eyes, and they make the same thing happen in their imagining – probably because this is what they assume the experimenter meant when he or she asked them to "scan" their image.

The empirical test of whether the pattern in such cases (including the mental color-mixing example cited earlier) is due to the architecture or to the representation is to ask whether the pattern can be systematically and rationally altered by changing beliefs about the task. That was the point of our experiment, which showed that this is indeed what happens. This shows that the pattern is what we call "cognitively penetrable" and allows us to conclude that it does not arise from a property of the cognitive architecture.

4 Unconscious Processes

What goes on in mental imagery, as well as what goes on in understanding linguistic utterances such as those discussed earlier, looks to be largely some kind of reasoning – drawing conclusions from what you know. But the nature of this reasoning is entirely opaque to your conscious experience. This is a universal finding: Most of what we need to hypothesize in order to explain how the mind works is *not* available to introspection, and what is available to introspection is usually not what is relevant – it's not what is doing the work. Recall our earlier example of assigning a referent to a pronoun in two sentences. This case clearly illustrates that the pronoun assignment depends on reasoning and on drawing inferences from facts that you know about the world, about social interactions, and perhaps even about the characters John and Fred if the sentences have occurred in a story. But you normally have no awareness whatsoever of there being any inferences involved in your understanding the sentences. There is rarely any consciousness of processing sentences. Yet it is known that complex grammatical analyses are involved. You need to uncover what is called the "thematic" or "logical" structure of the sentence – to discern who did what to whom. And this involves an extremely complex process known as "parsing," which entails a large number of rules of grammar, some specific to English, some relevant to all languages, and some idiosyncratic to a particular (discourse or story) context or to particular individuals. But you have no awareness of any of these.

We should view it as a major discovery of twentieth-century cognitive science that most of what goes on when we act intelligently is not available to conscious inspection. And since what goes on is reasoning, we have every reason to believe that it takes place in a system of symbols – that's the only way we know of doing it without hiding a homunculus inside the machine.

4.1 How can we know how it works inside?

Where do these observations leave a *science* of mind (e.g. cognitive science)? If you can't rely on introspection of your conscious experience to tell you what's going on in your mind, and if you can't rely on looking inside the skull using biological techniques to tell you what psychological processes are taking place, then how in the world *can* you tell? Of course, you can observe the organism in various ways in the laboratory. But if you are observing only the visible behavior – the input–output behavior – then can you distinguish among theories that produce the same input–output behavior? If the answer is no, then we are in trouble, because science is not interested in merely predicting input–output behavior. It is interested in the question: *how does it work*? And to say how it works is to do much more than predict what output it will produce, given a particular input.[5] At the very least, it is to specify the form in which representations are encoded and to give the algorithm by which the input–output function is computed in detail. But how can we do this if we do not have access to the program, if we cannot look inside the black box but are confined to examining only the organism's observable behavior?

This is a serious question and has been debated from time to time by both philosophers and psychologists. Yet strangely enough, experimental psychologists have been proposing and verifying detailed theories of how information is processed for the past 35 years. How can they do that? Here is the issue. If you say that all you have to go by is input–output behavior, you are making the methodological mistake underlying the ideology of behaviorism. You are not only assuming that all you have is a record of observed behavior, but also that any record of behavior is like any other record of behavior. Recall that we noted earlier that even the strict behaviorist must put up posters to solicit subjects, and when he does that, he assumes that it is the meaning of the sentences on the poster that is relevant to whether subjects will show up. Similarly, if the investigator is gathering data in an experiment and the subject says something like "Oops, I meant to hit the button on the right but hit the one on the left by mistake," no scientist, no matter how ideologically pure, will write down as part of the record, along with the list of buttons that were pressed, "Response 12: S uttered 'Oops, I meant . . .'" Rather, the scientist will do something like mark the response as an error, delete the erroneous response, start the trial over, or reject the subject's data. Why? Because some responses are taken to be the pure outputs of the system being studied, and some are taken to be statements about what the subject thought. A linguist gathers data by recording sentences in the language and examining patterns of co-occurences (e.g., what types of phrases go with what other types, and so on). But he or she also asks native speakers of the language such questions as whether a particular sentence (call it A) is a grammatically acceptable sentence, whether it is ambiguous, whether it means the same as another sentence B, or whether the meaning of sentence A is related to the meaning of sentence B as sentence D is related to sentence C (e.g., "John hit the ball" is to "The ball was hit by John" as "The dog chased the cat" is to "The cat was chased by the dog"; and "John hit the ball" is to "Who hit the ball?" or to "What did John hit?" as other related pairs of sentences you can easily think up. The linguist takes the answers to such questions to be not merely a sample of the language but as the speakers' judgements about the test sentences – as truthful claims *about* sentences. There is a world of difference between sentences that form the data-base of observations and sentences that constitute expert judgments about these sentences.

So here is one possible way to do better than merely trying to reproduce the input–output behavior that is observed in a laboratory. Ask the subject to tell you what he is trying to do, or what he is doing, or what he knows at various times during the process. This method (which of course applies only to more deliberate, conscious, and relatively slow processes, such as solving crossword puzzles or playing chess) has been used a great deal and is referred to as "protocol analysis." Protocol analysis is an instance of a more general class of methods for gathering evidence of intermediate states in the process. If your theory says that a certain input–output (I–O) behavior is the result of a certain program, then a way to test this is to ask what intermediate stages the program goes through – what partial solutions it has at various stages – and to compare this with the intermediate stages that a subject goes through. Such intermediate-state evidence can often be obtained by asking subjects to "think out loud" while solving a problem. But there are many other, more subtle ways of getting such data. For example, one can use eyetracking equipment to record what a subject is looking at (say, while doing an arithmetic problem or while reading). Such evidence tells you whether the subject is "looking ahead" or examining the problem in some unusual order. Scientists Alan Newell and Herbert Simon have used the protocol-analysis technique to great advantage in formulating and testing computational theories of problem-solving processes. More and more clever intermediate-state evidence has been accumulated by creative investigators.

Intermediate-state evidence, however, is not always available, especially for rapid and highly fluent processes such as visual perception. And it is sometimes misleading, since subjects can (and do) report what they think they were doing rather than what they actually were doing. But no technique is perfect by itself, and science always relies on converging evidence from many sources

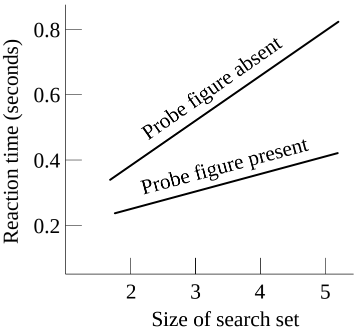

Figure 1.7 The graph on the left illustrates the results expected if search for the probe target was serial and self-terminating. The one on the right, which was actually found by Sternberg (1966), is what you would expect if the search was serial and exhaustive (did not stop when the target was found).

to increase the confidence level of its conclusions. Fortunately the arsenal of techniques for obtaining evidence of what process is being carried out is limited only by the creativity of the scientists, and every month new techniques and analysis tools appear in publications.

A major source of evidence in cognitive science is what I have called "relative complexity evidence." A major example of this entails the use of reaction times. If you vary some property of the task in a systematic way and observe an equally systematic change in the time taken to produce a response, you may be able to exclude one type of model and provide support for another. Probably the most frequently cited example of the use of this method is a study by Sternberg (1966). Sternberg did the following experiment. He asked subjects to memorize a small set of items – say, the letters G, A, F, and T (called the search set) – and then showed them one letter (called the probe) on a screen. Subjects had to press one button as fast as they could if the probe was a member of the memorized set and another button if it was not. The question Sternberg was asking is: How does a subject look up items in short-term memory? He found that it took longer to decide that the probe was not a member of the memorized set than to decide that it was. But that in itself tells us very little. What tells us much more is to examine how the reaction time increases as the size of the search set increases and to look at this function for positive cases (where the probe was a member of the search set) and for negative cases (where the probe was not a member of the search set). Figure 1.7 shows what Sternberg found (the graph and the numbers here are for illustrative purposes).

The relevant properties of these data are that the time it takes to find the probe in the search set increases linearly as the size of the set increases. This suggests that a serial search through the items in memory is taking place – *despite the fact that it seems to the subject to involve no processing and no search at all!* It looks from the figure as if each additional item adds about 0.07

seconds to the reaction time (i.e., the search time, or slope of the curve, is 70 milliseconds per item). But perhaps even more surprising is the comparison of the slopes of the case where the probe item is actually in the memory set with the case where it is not. We would expect the slope to be about twice as great when the probe is not in the memory set. Why? Because to establish that the item is not found in the memory set the subject must continue to search until the end of the list is reached. In that case, then on average twice as many items would have to be examined. For example, if the search set has four items and the probe is in the set, then on average (with the correct item being located in random positions in the set) it will be successfully located after two comparisons. But if it is not in the set, the subject will have to search through to the end to discover that, resulting in four comparisons. But Sterberg found that the slopes in the two cases were the same (as shown in the graph on the right in figure 1.7). So these results suggest that locating an item in a set that is stored in what is called short-term memory is accomplished by a *serial exhaustive search algorithm.* Although the facts are slightly more complicated than this, and disagreements still persist, this example illustrates the use of the what might be called *relative complexity methodology* to decide among different possible processes all of which could produce the same input–output behavior. In this case measurements of the relative time it takes for different task parameters helps to decide whether the lookup of the probe happens in parallel or serial, whether the search is exhaustive or self-terminating, and so on. Just knowing that people can tell whether a probe was a member of the search set is itself not very interesting, since any of a large number of possible processes could accomplish that.

Notice that we call this example an instance of the relative complexity methodology, because we are not actually interested in the amount of time it takes (since that depends on a lot of things, many of which, like the strength and mass of the fingers used to hit the button, have nothing to do with the information processing itself). Rather, time is simply being used as a measure of how many steps it takes. We assume that the more steps it takes, the longer will be the reaction time. But there are other ways to measure relative complexity. For example, if more errors are found under one condition than another, this could also be because the more error-prone condition requires more steps (on the assumption that there is more opportunity for an error to occur). Other more subtle measures are also possible and depend upon a more detailed mathematical model of the process. For example, subjects can decrease their response time by sacrificing accuracy. But this sacrifice, called the speed–accuracy trade-off, can itself have a different pattern depending on how difficult the response is (as opposed to how difficult the memory search is), and this can be analyzed mathematically to tell you whether a decrease in speed is due to more time being spent on the search or on the decision as to which response button to press. This is the beginning of a "stage analysis," which is a more refined way of breaking down the process into different stages, and for which some extremely refined techniques are available.

It is also possible to obtain evidence both for stages and for which of several possible processes are actually being used by the subject, by finding certain measurable properties that we have independent evidence to believe are correlated with different operations, or stages. For example, there are certain patterns in human brain waves (or EEGs) that are known to be correlated with detecting and recognizing a stimulus, as opposed to preparing to make a response. One of the more interesting of these so-called *event-related potential* patterns is known as the P300 pattern (because it consist of positive spikes occurring about 300 milliseconds after a stimulus). There is reason to believe that the duration of the P300 pattern may tell us how much time is spent recognizing the stimulus, even when the actual overt response takes longer and is affected by different variables. This technique has been used to investigate whether certain manipulations of the stimulus (say, making it dimmer or noisier or a rarer type – e.g., an infrequent word) that are known to slow down reaction time also slow down recognition time, or whether the slowdown occurs in the response-selection stage. Similarly, *galvanic skin response* (or "lie detector" response) can be used to show that a stimulus has been registered, even if the subject is unaware of it, and so on. There is no limit to the kinds of evidence that can be brought to bear on deciding what process is being used to derive a response. In each case the method depends on an assumption about what the measurement is actually related to in the information processing, just as we assumed that reaction time was related to number of operations. But the step from observation to theory always depends on such methodological assumptions, which have to be justified independently – and this is true in every science. In other words, there is nothing special about finding out how the process inside the black box works – even without opening it up and looking inside. It's the same as finding out what makes water change from liquid to solid at low temperatures. You don't do it by "looking inside." The secret, as elsewhere in science, is just to be clever!

4.2 What, then, is really in your mind?

The answer to the question "What's in your mind?" is that, although we don't know in any detail, we think that it will turn out to be symbolic expressions, and that thinking is some form of operation over these symbolic codes. The symbolic codes are likely to be quite different from any contemporary calculus or language (e.g., the symbols would have to be able to encode procedures as well as facts), and the operations over these symbols are likely to be very different from those encountered in contemporary computer languages, in probably being richer and possibly making use of apparatus that evolved for other purposes – like perception. But so far, nobody has been able to come up with anything that feels more natural – that looks like the objects of experience – and that is able to do the job. Moreover, whenever people have tried to propose radically different schemes – for example, ones that *look* like the nervous system – it has turned out to be the case that the *looking-like* was very superficial, and that in order to be able to reason, we still need to invent another layer of organization corresponding to some language-like combinatorial symbol system.

Of course, this all sounds strange and unnatural. But think of how unnatural the idea of a heliocentric planetary system, with planets kept in place by invisible forces acting at a distance, must have sounded to the Aristotelians of the seventeeth century, and how unnatural is the scientific answer to the question "What is in this table?" (i.e., almost entirely empty space with a variety of unobservable forces acting on unobservable particles and electromagnetic wave patterns). That's the way it is in science: Things are rarely what they seem. But over time, we all learn to live with the strangeness, and it usually becomes the mundane orthodoxy of the next generation.

Notes

[1] In this connection I recommend a paper by Bill Brewer (1974), in which he examines the vast literature on classical and operand conditioning of adult human subjects and finds in each case that the pattern of responses is best explained in terms of what the subject is led to *believe* about the outcomes of different voluntary actions.

[2] Here is an exercise you might perform for yourself. Suppose we are interested in how a person, as opposed to an electromechanical device, does arithmetic. Can you think of some empirical observations – along the lines of those implied by the questions above – that you could make that would help decide how the person carried out the arithmetic? For example, measuring the time taken to perform the task when the inputs are varied in certain systematic ways has been one of the main sources of evidence in cognitive science. Would some of the patterns of behavior tell you more about the biological level than about the computation carried out by the mind (the way it does in this example for certain of the observations listed in questions 1 and 2)? After you have thought about this for a while, you might look at the section "How can we know how it works inside?"

[3] See, however, the extended discussion of these issues in Pylyshyn 1984 and Fodor and Pylyshyn 1988.

[4] See the original study described in Kosslyn et al. 1978 as well as the subsequent discussion in Pylyshyn 1981.

[5] A theory that accounts only for the input–output behavior observed in a laboratory is said to be "weakly equivalent" to the subject being modeled. A theory that claims to tell you by what means (i.e., by what algorithm or program) the input–output behavior is generated is said to be "strongly equivalent" to the subject being modeled. Pylyshyn 1984 is mostly about what it takes to be a strongly equivalent theory.

References

Brewer, W. F. (1974) There is no convincing evidence for operand or classical conditioning in adult humans. In W. B. Weiner and D. S. Palermo (eds), *Cognition and the Symbolic Processes*, Hillsdale, NJ: Erlbaum 324–48.

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

Kosslyn, S. M., Ball, T. M. and Reiser, B. J. (1978) Visual images preserve metric spatial information: evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance* 4, 46–60.

Pylyshyn, Z. W. (1981) The imagery debate: analogue media versus tacit knowledge. *Psychological Review* 88, 16–45.

—— (1984) *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, MA: MIT Press.

Sternberg, S. (1966) High speed scanning in human memory. *Science* 153, 652–4.