It's hard enough arguing that the research program of AI will eventually succeed. But it seems to many people that it could succeed completely and still not provide a theory of *phenomenal consciousness*. I use the term "phenomenal" to make it clear that I am not talking about other concepts of consciousness. There are several other kinds to check off:

- Not being asleep: You are conscious in this sense when you aren't unconscious.
- Attentiveness: You can be unaware that there is a high-pitched hum in your vicinity until someone points it out.
- Accessibility to report: You are unconscious of what your brain does when it processes visual information; you are conscious of what you do when you are looking for your glasses, but unconscious of what makes you more attuned to glasses-shaped objects when you're looking.
- consciousness as being aware of oneself: In playing chess you might be conscious of your sweaty hands and pounding heart, or you might be focused entirely on the game, and hence "unconscious" or at least "unselfconscious."

All of these are of interest, but none of them constitutes the "hard problem" of consciousnesss, to use David Chalmers's phrase. The hard problem is that red things look a certain way to me, different from green. I might be able to build a computer that could distinguish red things from green ones, but, at first glance, it doesn't seem as if either color would look "a certain way" to the computer. The way things are experienced by conscious beings are called the *qualia* of those things, and explaining what it is to have qualia is the hard problem of consciousness. When it's

important to focus our attention on this meaning of the word "conscious," I will use the term "phenomenal consciousness."

As I said, it looks at first as if computational models could never explain phenomenal consciousness. They just don't trade in the correct currency. They talk about inputs, outputs, and computations, but sensations of any of these things don't enter in. Hence it seems logically possible that there could be a computer that made exactly the same sensory discriminations as a person, and yet experienced nothing at all. What I would like to argue in this chapter is that, if there is ever such a thing as an intelligent robot, then it will have to exhibit something very much like consciousness. I will then take the further step of arguing that this something is exactly what we mean by human consciousness, in spite of our intuitions to the contrary.

Before I make those arguments, let me remind you of the main argument of chapters 1 and 2. As long as the brain remains poorly understood, there will always be room to assume that some noncomputational essence within it makes consciousness happen. However, it may not remain that poorly understood much longer. We can already map the nervous systems of very simple creatures (leeches with only a few dozen neurons), and we get the same intuitions as when we look at computers: we can *see* what is happening in the nervous system, we can model it perfectly well as a kind of computation, and we can *see* the absence of experience. So to maintain a belief in dualism we have to believe that the human brain contains structures that are quite different from those in leeches, structures that would cause experience to happen. The problem is that we have no idea what those structures might be. Worse, to the extent we do understand what's going on in the brains of humans and other mammals, the events don't seem to be qualitatively different from what goes on in the "brains" of leeches.

Sooner or later, I predict, we're going to be faced with trying to explain consciousness without resort to any structures or mechanisms that are significantly different from the ones we now understand. If you think this is preposterous, then you may not be able to follow the argument much further. On the other hand, if you find unsatisfactory all the other proposals that have been made (Penrose 1994; Chalmers 1996; O'Brien and Opie 1999), then perhaps you will bear with me. I will warn you, however, that a computationalist explanation of consciousness will inevitably sound

like "explaining away" rather than true explanation. Almost any materialist explanation, even the correct one, is going to have this problem or a similar one, because of the wide gulf between our intuitions about matter and our intuitions about mind. In the end the correct theory will win the argument only if the evidence in its favor outweighs intuition. I can't claim to provide such evidence, but I can say what I think it will look like. You must judge whether it has the potential to trump some of your "undoubtable" intuitions.

The full explanation of consciousness in terms of computation will require a fairly elaborate argument. But some of the apparent puzzles of consciousness dissolve immediately when we adopt a computational perspective. For example, consider the classic issue of distinguishing visual images from actual visual experiences of real objects. As Armstrong (1968, p. 291) puts it, "It seems clear ... that there is the closest resemblance between perceptions ... and mental images. A good way to begin an inquiry into the nature of mental images, therefore, is by asking, 'What are the marks of distinction between perceptions and visual images?'" He then discusses proposals such as Hume's that mental images are somehow less vivid than real perceptions and that that's how we tell them apart.

If the brain is nothing more than a computer, this problem simply evaporates. It's like asking how an income-tax program tells the difference between dividend income and income from tax-free bonds. They're both numbers; they might even on some occasions be exactly the same number; how does the program tell them apart? The question is silly. Either they are never examined by the same process at the same point in the computation, or each is accompanied by a further bit of information that just *says* what category it falls into.

Or consider the problem Jackendoff (1987, p. 12) calls "the *externalization* of experience—the fact that my experience may be *of* things external to me.... The blueness of the sky is *out there in the sky;* the pain is *in my toe*.... [A materialist theory] claims that the experienced blueness *in the sky* is identical with a state of neurons *in my brain* and that the experienced pain *in my toe* is identical with another state of neurons *in my brain.* How can the same thing be in two different places?"

This is obviously a pseudoproblem if the brain is a computer. When a signal arrives from the toe, its content contains a specification of the

location of the pain, not just the fact that it is a pain. Experience arises as an aspect of the way these messages are processed; it is not a separate process in which we become aware of the existence and nature of the message itself. The message says where the pain is; otherwise there is no way in the world the brain could know where it comes from. That's why a lesion in the nervous system can cause pain to be experienced in a place far from the lesion. The lesion causes bogus messages to be sent, and their content is to some extent arbitrary; it would be a mere coincidence if the message happened to mention the exact location where the lesion is.[1]

Of course, none of this explains what subjective experience actually is. But it does clear away a whole class of problems that have vexed philosophers. Perhaps the others will also succumb to a computational explanation.

Free Will

I will start the explanation with a little warmup, explaining the phenomenon of free will. Many people have thought that free will has something to do with phenomenal consciousness. I actually don't think that, but they do have explanations that are similar in form.

Suppose we have a robot that models the world temporally and uses its model to predict what will happen. I am not talking about "mental models" as the term is used in psychology (Johnson-Laird 1983), but about numerical or qualitative causal models as used in simulation. Such models are a familiar application of computers. The main difference between what computers normally do and what my hypothesized robot does is that the robot is modeling the situation it is now actually in. This model includes various symbols, including one I'll call R, which it uses to denote itself. I dealt with the idea that computers manipulate symbols in chapter 2, and will discuss it at greater length later, especially in chapter 5. When I say the symbol denotes the robot itself, I don't mean to imply that the word "itself" implies something about "self." All I mean is that, for example, when it detects an object in its environment, it notes that R knows the object is present; and when it has a tentative course of action on hand, that is, a series of orders to be transmitted to its effector motors,

it will base its modeling activity on the assumption that R will be carrying out those actions.

Now suppose that the actual situation is that the robot is standing next to a bomb with a lit fuse. And suppose that the robot knows all this, so that in its model R is standing next to B, a bomb with a lit fuse. The model is accurate enough that it will predict that B will explode. Supposing that the robot has no actions on its agenda that would make it move, the model will predict that R will be destroyed.

Well, actually it can't make this prediction with certainty, because R will be destroyed only if it doesn't roll away quickly. The conclusion that it would not roll away was based on the robot's own current projection of what it is going to do. But such projections are subject to change. For instance, the robot might be waiting for orders from its owner; a new order would make it roll away. More interestingly, the robot might have a standing order to avoid damage. Whenever its model predicts that it is going to be damaged, it should discard its current action list and replace it with actions that will protect it, assuming it can find some. Finding actions to achieve goals is a deep and fascinating topic, but it needn't concern us here. The robot concludes it should exit the room, and does so.

What I want to call attention to is how this sequence of events is represented in the robot's model, and how that will have to differ from reality. The reality is that the robot's actions are entirely caused by events. The sequence I laid out is a straightforward causal chain, from perception, to tentative prediction, to action revision. But this causal chain cannot be represented accurately in the model, because a key step of the chain, the making of tentative predictions, involves the model itself. The model could not capture this causal chain because then it would have to include a complete model of itself, which is incoherent. In other words, some of the causal antecedents of R's behavior *are situated in the very causal-analysis box* that is trying to analyze them. The robot might believe that R is a robot, and hence that a good way to predict R's behavior is to simulate it on a faster CPU, but this strategy will be in vain, because this particular robot is itself. No matter how fast it simulates R, at some point it will reach the point where R looks for a faster CPU, and it won't be able to do that simulation fast enough. Or it might try inspecting R's listing, but eventually it will come to the part of the listing that says "inspect R's

listing." The strongest conclusion it can reach is that "If R doesn't roll away, it will be destroyed; if it does roll away, it won't be." And then of course this conclusion causes the robot to roll away.

Hence the robot must model itself in a different way from other objects. Their behavior may be modeled as caused, but its own (i.e., R's) must be modeled as "open," or "still being solved for." The symbol R must be marked as exempt from causal laws when predictions are being made about the actions it will take. The word "must" here is just the "must" of rational design. It would be pointless to use a modeling system for control of behavior that didn't make this distinction; and it would be unlikely for evolution to produce one.

Any system that models its own behavior, uses the output of the model to select among actions, and has beliefs about its own decisions, will believe that its decisions are undetermined. What I would like to claim is that this is what free will comes down to:

A system has free will if and only if it makes decisions based on causal models in which the symbols denoting itself are marked as exempt from causality.

By this definition, people have free will, and probably so do many mammals. There are probably many borderline cases, in which an animal has a rudimentary causal model, but the exemption from causality is given to its self symbols by building in some kind of blind spot, so that the question can't come up, rather than by providing a belief system in which there are peculiar beliefs involving the self symbol.

People lose their freedom when they cease to believe that their decisions depend on their deliberations. If you fall out of an airplane without a parachute, you may debate all you like about whether to go down or up, but you know your deliberation has no effects. More subtly, an alcoholic or drug addict may go through the motions of deciding whether to indulge in his vice, but he doesn't really believe the decision is a real one. "What's the use," he might think, "I've decided every other morning to have a drink; I know I'm just going to make the same decision; I might as well have one." In this case the belief in one's own impotence might be delusional, but it's self-fulfilling. One can contrast the addict's situation with the decision of whether to take a breath. You can postpone breathing only so long; at some point the question whether to breathe or not seems to be "taken out of your hands." The alcoholic classes his decision to take a drink as similar to a decision to take a breath. He no longer believes in his freedom, and he has thereby lost it.

The obvious objection to this account is that it declares a certain natural phenomenon to be free will, when introspection seems to proclaim that, whatever free will is, it isn't *that*. It appears to identify free will with a *belief* in free will, and surely the two can't be the same. It's as if I declared that divinity is a belief that one is God, so that any schizophrenic who thought he was God would be divine. This objection might have some weight if I actually did identify free will with a belief in free will, but I don't. Rather, I identify free will with a belief in exemption from causal laws. The alternative formulation is not just implausible, but vacuous.

Still, this identification of free will with a certain computational property may seem disappointingly trivial. It has nothing to do with autonomy, morality, or the worth of the individual, at least not at first glance. I admit all this. Unfortunately, this is the only concept of free will the universe is likely to provide. Many volumes have been written about how freedom might find a place in a world subject to physical laws, and no one has ever succeeded in explaining what that might mean. Some find comfort in the indeterminacy of quantum mechanics, or even in the lack of predictability in the classical laws of physics, but freedom surely doesn't mean randomness. Some suppose that free decisions are those that "might have been otherwise," but it is notoriously difficult to say what this means. So we are in the odd position of being introspectively certain of something that makes no sense.

In such a situation, the problem should shift to explaining why we have that introspective view, not how it might actually be true after all. Once we make that shift, the problem resolves very simply, along the lines I have indicated.

Some may find this to be a scary tactic, with implications that may be hard to control. What else are we going to throw overboard as we proceed? Suppose we show that moral intuitions are incoherent. Do we then simply shift to explaining why we have moral intuitions? Does that mean that we need not be bound by moral intuitions?

I admit to finding this scary myself, but the case of free will gives us a bit of reassurance. Even though I accept my account of free will, it doesn't change the way I think about my actions. As many philosophers

(notably William James) have pointed out, it makes no sense to order one's life as though we could not make free choices. A statement of the form, "Because we can't make decisions, we should..." is silly, because any statement about what we "should" do presupposes that we can make decisions. We're stuck with free will.

This dismissal of qualms is more glib than I intend. I will come back to this topic later (chapter 6). But first, let us see if the method used to explain free will will also explain, or explain away, qualia.

Let's look more carefully at the structure of deliberate choices. Suppose a robot is built to sense and avoid extreme heat. One way to make it avoid heat is to build in reflexes, analogous to those that cause animals to jerk back suddenly when they touch something hot. But a reflex won't get the robot out of a burning building. That requires planning and executing a long string of actions. The detection of heat, and the prediction that it's going to get worse, must cause the robot to have an urgent goal of getting out of the situation it's in. A *goal* in artificial-intelligence terminology is a structure describing a state of affairs a system is to try to bring about. In this case the goal is "that R [the robot] be out of the building." The robot might have other goals, such as "that R know whether any human is in the building."

The interaction between goals can be complicated. Suppose the robot's second goal (call it G2) was an order given when there was no reason to believe that there were any persons left in the building. Then the robot's owners might prefer that it save itself (goal G1) rather than continue to search for people who probably don't exist. On the other hand, if fewer people can be accounted for outside the building than were believed to have been inside, then G2 would take precedence. But even though the robot decides not to flee the building right away, this is not a decision it can just make and forget. As the heat becomes more intense, the probability that it will be destroyed increases. There may come a point when wasting a perfectly good robot for a negligible chance of saving humans may seem foolish. I am not assuming that every intelligent creature must have an innate and overriding desire for self-preservation. It should be possible to build a robot with no desire at all to preserve itself. But we may as well imagine that the builders of the robot put in a desire to avoid destruction of the robot just so it would allow its own destruction only when its owners wanted it to.

It may sound odd to require a robot to have a "desire" for something. If we want the robot to behave in such a way as to bring an outcome about, why can't we just program it to do that? Isn't talking of "desires" just quaint anthropomorphism? No, it isn't. As I explained above, a robot whose world model is rich enough to include itself must believe itself to be exempt from causality. That means that if we want to program a behavior into such a robot we must arrange for the robot to believe that there is a good reason to choose that behavior. In other words, there has to be something that looks like evidence in favor of one course of action compared with the other. The robot may believe that everything it does is caused, but it will still have to have reasons for its choices. I said in chapter 1 that the distinction between reasons and causes suggests an argument in favor of dualism. Now we see that even robots must make this distinction and might thereby be tempted to be dualists themselves.

To make the point vivid, suppose that in the course of fighting its way through the burning house the robot, call it M, encounters another robot with an entirely different set of goals. The other robot, call it C, might be the intelligent controller of the house in question, and it might have been instructed to burn itself up. (The house has been condemned; there's no further need for this system to want to preserve itself.) C inquires of M why it is moving so steadily but hesitantly toward a burning room in which a baby is located. M replies that it wants to save the baby but doesn't want to be destroyed. C might ask, "Why are you taking corridor A to that room instead of corridor B?" And M might respond "There is less fire in corridor A." And so forth. But if C asks, "Why don't you want to be destroyed?" M will run out of answers. One answer might be, "Because I was designed that way." But M may not know this answer, and in any case it is an answer about what *causes* his goal. C wants to know his reason for having the goal. There is no reason for this goal; it is its own reason. M believes that "my destruction would be bad" is true and self-evidently true.

The point is that reasons must come to an end. It's conceivable that the end could be located somewhere else. The robot could believe that the happiness (or, as decision theorists say, the "utility") of its owners is the highest purpose, and thus want to preserve itself only as long as it believes that preserving itself is likely to cause its owners more happiness than the

alternatives in the present circumstances. But it's hard to visualize this scheme working. Most of the time the robot cannot judge all the factors contributing to the happiness of its owners, or how its preservation would affect them. It's going to be more practical to have it want self-preservation unless that directly contradicts an order from the owner.

Now suppose robot M is in the burning room. Its temperature sensors are going off their ranges. The search has so far not revealed the location of the baby, if indeed it was ever here. The urgency of goal G1 is getting higher and higher, the likelihood of achieving G2 getting lower and lower. Eventually the robot decides to give up and run.

Until that moment, there is a "detachment" between the output of a perceptual module and the way it is used. When the temperature sensors report "Extreme heat," a goal is set up to flee, but it might not be acted on. Even so, the sensor report is impossible to ignore. Even if it doesn't get any worse, it is constantly demanding attention, or, more precisely, demanding computational resources to evaluate whether it is necessary to act on it. As long as the robot decides to stay in the fire, the heat is labeled as "unpleasant but bearable." At this point we can conclude that the robot's perception of the fire has something like a quale of unpleasantness. I do not mean that the robot labels the state reported by the sensor with the English word "unpleasant." I mean that however the state is represented, it is classified as "to be avoided or fled from," and it is so classified *intrinsically.* Just as a chain of goals must come to an end with a goal that can't be questioned, so must evaluations of sensory states. The robot may dislike going into burning buildings because it dislikes heat. But it doesn't dislike heat because of some further bad consequences; high heat is intrinsically not-likable. As with the goal of self-preservation, we can easily imagine the chain continuing. The robot might not like sensing heat because it is likely to lead to a state where it will sense damage. But the chain has to stop somewhere, and the sensing of extreme heat is as good a place as any. Extreme heat is easier to detect than damage and is strongly correlated with it.

You may balk at the notion that I have actually explained a quale, and I was careful to use the phrase "something like a quale." For one thing, at most I have explained one dimension of an experience, the dimension of "pleasantness." Wine and cheese may each taste pleasant, but they

differ in lots of other ways. One may in fact doubt whether pleasantness is part of the quale at all. It seems clear to me that it is, and that the difference between the taste of turkey before Thanksgiving dinner and the taste afterward is explained by a difference in pleasantness (Dennett 1991). But it's not clear that a sensation could consist in pure pleasure or pure pain with no other characteristics. So we haven't yet endowed our robot with even an "as-if" sensation.

Still, we have given it an important component of mental life, namely *preferences,* which seem closely allied, at least in people, with emotions. With conflicting goals, a creature must have tags giving the relative values of various situations, and there is no point in having the values be questionable. If something can be questioned, then there must be a way of weighing pros and cons, and the factors in that weighing must be unquestionable. A creature that could really question the value of everything would never act.

A creature without preferences can behave, but it cannot make deliberate choices. Nowadays air-to-air missiles are programmed to avoid heading toward an airplane with a "Friend or Foe" signal that identifies it as a friend. An intelligent attack robot might want to be able to entertain the hypothesis that a friendly aircraft had been captured by the enemy. It would have to weigh its repulsion away from the possibility of attacking an aircraft labeled friendly against its attaction toward the possibility of destroying the enemy pilot sitting in that aircraft.

In science fiction, robots and androids are often portrayed as being without "emotion." In a typical plot, an android will be portrayed as unable to love or laugh (until a special experimental chip is added). It is, however, able to carry on a conversation, have multiple goals, and decide on different courses of action. It often prefers deduction to induction, and is usually driven by curiosity. In other words, it is not without preferences, it just prefers different things than the average human does. If you ask it, why do you spend time trying to find out about humans instead of studying more mathematics, it will give answers like "Humans have always fascinated me." If you ask it, why do you help the Rebel Alliance and not the Evil Empire, it will give answers like, "I find the Emperor and his minions suboptimal," as if robots, as ultrarational beings, would have an inherent tendency to try to make situations optimal, without actually

preferring anything. One might dismiss all this as sloppy and inconsistent imagining by whoever wrote the script. But try to imagine an android that really had no preferences at all. It would behave in such a way as to bring certain goals about, as monomaniacally as a pool pump behaves when it keeps water circulating in a swimming pool. But when asked it would never admit to having any preferences for one outcome over another. When asked, "Why did you steer the ship left instead of right?" it would answer, "There was no reason for what I did. If you inspect my program you can see that the cause of my behavior is a long string of computations which I will print out if you wish." The problem is that it can't extend this way of thinking into talking about the future. If you ask, "Should we go left or right?" the android will refuse to answer the question on the grounds that it has no preference one way or the other. You have to ask it, "Given the following criteria, should we go left or right?" and then spell them out.

You might suppose that you could tell the robot, "Adopt the following criteria until I countermand them," but that just means imagining the original android again. It doesn't adopt the criteria for any reason ("You may read my program..."), and once they are adopted they *become* its unquestionable reasons for further decisions. There is no difference between an android that really prefers X to Y and one that unquestioningly adopts a preference for X over Y when told to.

Okay, but the science-fiction author never said the android didn't have preferences. She said that it didn't have *emotions*. It's interesting that to convey this fact the writer has the android behave as a human would if the human were heavily sedated or in shock. As long as the android doesn't have emotions, why not have it chuckle occasionally just to brighten the days of the people around it?

The question is, however, whether there can be preferences without emotions. Emotions seem to have three components: a belief, a preference, and a quale peculiar to each emotion. *Fear* is a belief that something is likely to happen, a preference that it not, and a set of sensations peculiar to fear. *Regret* is the belief that something has already happened, a preference that it hadn't, and a different set of peculiar sensations. (Obviously, there are many nuances here I am neglecting.) So it seems logically possible that one could have a preference and a belief without any special

sensation. To investigate this further we have to focus on the structure of perception.

Modeling Perception and Judgment

Once again let's imagine the case of a robot, only now what the robot is thinking about is perception, not action. The robot has just made a perceptual mistake. It saw a straight object that it took to be bent. It stuck a stick into a pool of water and observed the stick change chape. However, after doing various experiments, such as feeling the object as it entered the water, it decides that the stick never actually bends, it just appears to.

This story sounds plausible, because we've all experienced it ourselves, one way or another. Actually, there is no robot today that could go through this sequence of events, for several reasons. First, computer vision is not good enough to extract arbitrary, possibly surprising information from a scene. A typical vision system, if pointed at a stick in a tub of water, would probably misinterpret the highlights reflected from the water surface and fail to realize that it was looking at a tub of water with a stick thrust into it. Assuming it didn't stumble there, and assuming it was programmed to look for sticks, it might fit a line to the stick boundary and get a straight stick whose orientation was halfway between the orientation above the water level and the orientation below. Or it might see one half of the stick, or two sticks.

Even if we look forward to a time when computer vision systems work a lot better than they do now, there are still some gaps. There has been very little work on "cross-modality sensor fusion," which in this case means the ability to combine information from multiple senses to get an overall hypothesis about what's out there. No robot now is dexterous enough to feel the shape of a stick, but even if it were there would still be the problem of combining the input from that module with the input from the vision module to get an overall hypothesis. The combination method has to be clever enough to know when to reject one piece of information completely; taking some kind of average of the output of each sense will not be useful in the case of the unbent stick.

Even if we assume this problem can be solved, we still don't have the scenario we want. Suppose the robot is reporting what it senses. It types

out reports like this:

```
Stick above water
Stick goes into water; stick bent
(Feels)
Stick straight
```

The question is, How does this output differ from the case where the stick was really bent, then straightened out? For some applications the difference may not matter. Suppose the robot is exploring another planet, and sending back reports. The humans interpreting the output can realize that the stick probably didn't bend, but was straight all along. Let's suppose, however, that the robot actually makes the correct inference, and its report are more like

```
Stick above water
Stick goes into water; stick bent
(Feels)
Correction: stick never bent
```

We're still not there; we still don't have the entire scenario. The robot isn't in a position to say that the stick *appeared to be* bent. Two elements are missing: The first is that the robot may not remember that it thought the stick was bent. For all we know, the robot *forgets* its earlier report as soon as it makes its new one. That's easy to fix, at least in our thought experiment; as long as we're going far beyond the state of the art in artificial intelligence, let's assume that the robot remembers its previous reports. That leaves the other element, which is the ability to perceive the output of sensory systems. As far as the robot is concerned, the fact that it reported that the stick was bent is an unexplained brute fact. It can't yet say that the stick "appeared to be" anything. It can say, "I concluded *bent* and then I concluded *straight*, rejecting the earlier conclusion." That's all.

This may seem puzzling, because we think the terms in which we reason about our perceptions are natural and inevitable. Some perceptual events are accessible to consciousness, while others are not, because of the very nature of those events. But the boundary between the two is really quite arbitrary. For instance, I can tell you when something looks three-dimensional, whether it is or not. I know when I look through a stereoscope that I'm really looking at two slightly different two-dimensional

objects; but what I see "looks" three-dimensional. If someone were paying me money to distinguish between 3-D and 2-D objects, I would disregard the strong percept and go for the money. Another thing I know about stereo vision is that it involves matching up tiny pieces of the image from the left eye with corresponding pieces of the image from the right eye, and seeing how much they are shifted compared with other corresponding pieces. This is the process of finding *correspondences* (the matching) and *disparities* (the shifts). But I am completely unaware of this process. Why should the line be drawn this way? There are many different ways to draw it. Here are three of them:

1. I could be aware of the correspondences and disparities, plus the inference (the depths of each piece of the image) that I draw from it. In the case of the stereoscope I might continue to perceive the disparities, but refuse to draw the inference of depth and decide that the object is really 2-D.
2. I could be aware of the depths, but, in the case of the stereoscope, decide the objects is 2-D. (This is the way we're actually built.)
3. I could be unaware of the depth and aware only of the overall inference, that I'm looking at a 2-D object consisting of two similar pictures.

It's hard to imagine what possibilities 1 and 3 would be like, but that doesn't mean they're impossible. Indeed, it might be easier to build a robot resembling 3 than to build one resembling us.

Nature provides us with examples. There are fish called "archer fish" that eat insects they knock into the water by shooting them with droplets. These fish constantly look through an air-water boundary of the kind we find distorting. It is doubtful that the fish find it so; evolution has no doubt simply built the correction into their visual systems. I would guess that fish are not conscious; but if there were a conscious race of beings that had had to judge shapes and distances of objects through an air-water boundary throughout their evolutionary history, I assume their perceptual systems would simply correct for the distortion, so that they could not become aware of it.

The difference between people and fish when it comes to perception is that we have access to the outputs of perceptual systems that we don't believe. The reason for this is fairly clear: Our brains have more general mechanisms for making sense of data than fish have. The fish's brain

is simple, cheap, and "designed" to find the best hypothesis from the usual inputs using standard methods. If it makes a mistake, there's always another bug to catch (and, if worse comes to worst, another fish to catch it). People's brains are complex, expensive, and "designed" to find the best hypothesis from all available inputs (possibly only after consultation with other people). The fact that a perceptual module gave a false reading is itself an input that might be useful. The next time the brain sees that kind of false reading, it may be able to predict what the truth is right away.

Hence a key step in the evolution of intelligence is the ability to "detach" modules from the normal flow of data processing. The brain reacts to the output of a module in two ways: as information about the world, and as information about that module. We can call the former *normal access* and the latter *introspective access* to the module. For a robot to be able to report that the stick appeared to be bent, it must have introspective access to its visual-perception module.

So far I have used the phrase "aware of" to describe access to percepts such as the true and apparent shape of a stick. This phrase is dangerous, because it seems to beg the question of phenomenal consciousness. I need a phrase to use when I mean to say that a robot has "access to" a representation, without any presupposition that the access involves phenomenal consciousness. The phrase I adopt is "cognizant of."[2]

There is another tricky issue that arises in connection with the concept of cognizance, and that is *who* exactly is cognizant. If I say that a person is not cognizant of the disparities between left and right eye, it is obvious what I mean. But in talking about a robot with stereo vision, I have to distinguish in a non-question-begging way between the ability of the robot's vision system to react to disparities and the ability of the *robot* to react to them. What do I mean by "robot" over and above its vision system, its motion-planning system, its chess-playing system, and its other modules? This is an issue that will occupy us for much of the rest of this book. For now, I'm going to use a slightly dubious trick, and assume that whatever one might mean by "the robot as a whole," it's that entity that we're talking to when we talk to the robot. This assumption makes sense only if we *can* talk to the robot.

I say this trick is dubious for several reasons. One is that in the previous chapter I admitted that we are far from possessing a computational

theory of natural language. By using language as a standard part of my illustrations, I give the impression of a huge gap in the theory of robot consciousness. I risk giving this impression because it is accurate; there are several huge gaps in the theory, and we might as well face up to them as we go. Another risk in bringing language in is that I might be taken as saying that without language a system cannot be conscious, which immediately makes animals, infants, and maybe even stroke victims unconscious. In the long run we will have to make this dependence on language go away. However, I don't think the linkage between language and consciousness is unimportant. The fact that what we are conscious of and what we can talk about are so close to being identical is in urgent need of explanation. I will return to this topic later in the chapter.

Let's continue to explore the notion of cognizance. In movies such as *Westworld* and *Terminator,* the director usually feels the need, when showing a scene from a killer robot's point of view, to convey this unusual perspective by altering the picture somehow. In *Westworld* this was accomplished by showing a low-resolution digital image with big fat pixels; in *Terminator* there were glowing green characters running down the side of the screen showing various ancillary information. In figure 3.1 I have made up my own hypothetical landscape adorned with both sorts of enhancements; you may imagine a thrilling epic in which a maniacal robot is out to annihilate trees. What's absurd about these conventions is the idea that vision is a lot like looking at a display. The visual system delivers the information to some internal TV monitor, and the "minds's eye" then looks at it. If this device were used in showing human characters' points of view, the screen would show two upside-down images, each consisting of an array of irregular pixels, representing the images on the backs of their retinas. The reason we see an ordinary scene when the movie shows a person's point of view is that what people are normally cognizant of is what's there. The same would, presumably, be true for a killer robot.

What's interesting is the degree to which people can *become* cognizant of the pictorial properties of the visual field. Empiricist psychologists of the nineteenth century often assumed that the mind had to *figure out* the correct sizes and shapes of objects starting from the distorted, inverted images on retinas. A young child, seeing two women of the same height, one further away, would assume the further one was smaller (and upside

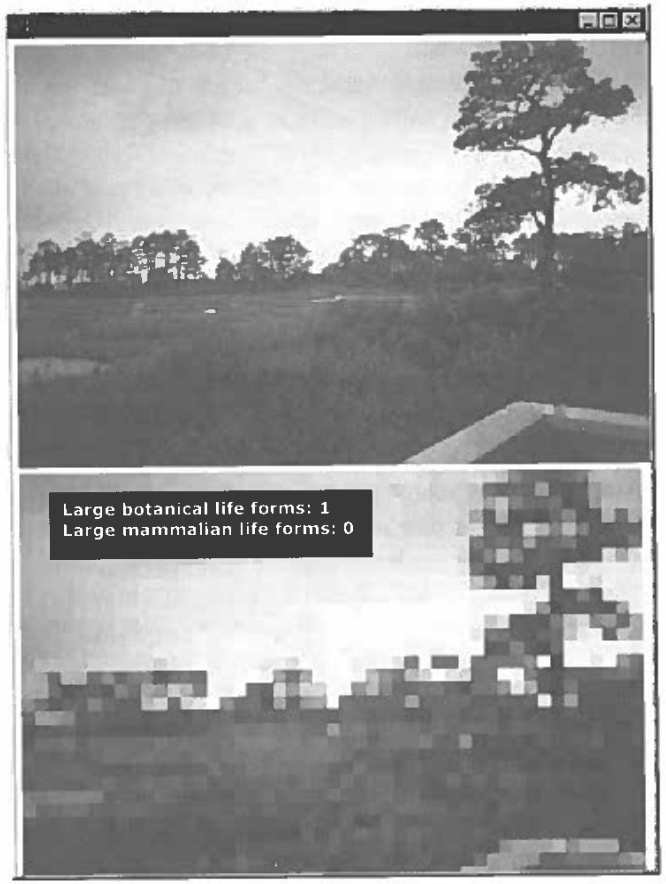

Figure 3.1
How Hollywood imagines a robot's visual experience

down), because her image was. He would eventually learn the truth somehow, that is, learn to infer the correct size and orientation of an object, and then forget he was making this inference; indeed, he would find it hard to become aware of it. Nowadays we know that the perception of the sizes of objects at different distances is an innate ability (Baillargeon et al. 1985; Banks and Salapatek 1983). What's remarkable, in fact, is that with training people can actually become cognizant of the apparent sizes of images on the retina.[3] This is a skill artists have to acquire in

order to draw images that look like actual images instead of schematic diagrams. Not everyone can do it well, but apparently anyone can grasp the idea. Look at two objects that are about the same size, two people, for instance, who are at different distances from your eye. Mentally draw two horizontal lines across the scene, one touching the top of the nearby object, the other the bottom. The faraway object can fit comfortably between the two lines with space to spare. If it doesn't, move your head up or down until it does, or find a different faraway object.

I could draw a picture of this to make it easier to visualize, but that would defeat the point I'm trying to make, which is that the average person, with training and practice, can view his or her visual field *as if it were a picture*. In doing this operation you are using the space of visual appearances in a way that is quite different from the way it is normally used. It is easy to imagine a race of beings with vision as good as ours who are incapable of carrying these operations out. They might simply be unable to see the visual field as an object at all. (Most animals presumably can't.) Or they might be able to draw imaginary lines across their visual field, but might be able to conceive of them only as lying in three-dimensional space. Asked to draw a horizontal line in the direction of a faraway person, touching the head of a nearby person, they might invariably imagine it as touching the head of the faraway person, as a horizontal line in space actually would. It is reasonable to suppose that there is some evolutionary advantage in having the kind of access that we have, and not the kind this hypothetical race would have.

Note that current vision systems, to the extent that they're cognizant of anything, are not cognizant of the two-dimensional qualities of the images they manipulate. They extract information from two-dimensional arrays of numbers (see chapter 2), but having passed it on, they discard the arrays. It would be possible in principle to preserve this feature of the introspective abilities of robots, even if they became very intelligent. That is, they could use visual information to extract information very reliably from the environment, but would never be able to think of the image itself as an object accessible to examination.

So far, so good; we have begun to talk about the way things appear to a perceptual system, but we still haven't explained phenomenal consciousness. That appears only in connection with certain kinds of introspection, to which we now turn.

Qualia

I argued above that a robot must assign values to different sensor inputs, but that's not the same thing as "feeling" them differently. We can imagine a robot attaching a number between −10 and 10 to every input, so that a dose of radiation and a fire might both get −9, but there must be more to it, or the robot wouldn't distinguish the two at all. You might have two pains, one shooting and one throbbing, that were equally unpleasant, but they wouldn't feel the same.

Of course no robot has a problem distinguishing one sensory input from another. The robot, we suppose, has several different sensors, and their reports do not get mixed up. A signal coming from the vision system does not get confused with a signal coming from the auditory system. Within a given sensory system there is similarly little possibility of confusion. A high-pitched sound yields one signal, a low-pitched sound another. Nonetheless, the question of how we distinguish a high-pitched sound from a low-pitched one can cause confusion. We're asking the *reason* for a judgment, and, as in the case of asking for the reasons for a decision, it is easy to mix this up with a request for the *cause* of the judgment. The cause is neurological (or computational): A physical transducer converts sound vibrations into signals, and low and high pitches yield different signals, which can be compared by another subsystem. But that's not the reason for the judgment. What I mean by "reason for a judgment" is exemplified by the case of distinguishing a fake Rembrandt from a real one. Here there is a list of aspects of the two objects that cause an expert to have an opinion one way or the other. In the case of high vs. low pitch, there are no such aspects, and hence no *reason* for the judgment (just as the robot has no *reason* to prefer surviving). Nonetheless people seem to have a reason where robots do not, to wit: "They sound different!"

Let's look closely at the sense system that has exercised philosophers the most, color vision. Let's start by supposing that a robot reacts to colors in a way isomorphic to ours. That is, its vision system is implemented using a system of three color filters sensitive in the same ranges as ours (Clark 1993), implemented with our visual pigments, or in some equivalent way. The robot cannot in principle make color judgments any finer than ours. That is, shown two different mixtures of light frequencies that looked

identical to people, it would classify them as identical also. We may also suppose that its judgments are not coarser than ours; it uses just as much information as we do. Further, let's assume that it can make judgments about the similarity of colors that are indistinguishable from a human's. Of course, different people judge such similarities differently, so the robot only has to make judgments that are in the same neighborhood as people's. We will collect a record of the robot's judgments by simply asking it which objects seem to have identical or similar colors, thus making use of my postulate that linguistic access is an accurate measure of cognizance.

Having given the robot the same powers of discrimination that people have, we now stipulate that these judgments of identity and similarity are *all* that the robot is cognizant of. Just as people have no introspective access to the fact that their color judgments are based on the differential sensitivity of three visual pigments, the robot has no access to the equivalent fact about itself (figure 3.2).

We now have the robot making humanlike judgments about colors. The next step is to get the robot to associate words with colors the way people do, either by training it or programming it. Since words match closely with similarity judgments, this shouldn't be hard. The word "green" will

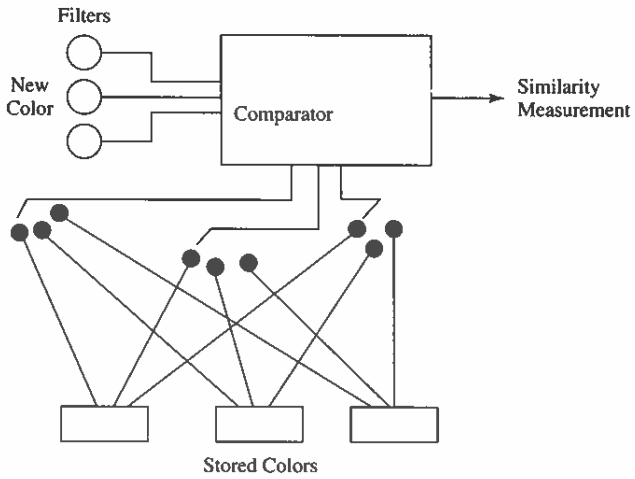

Figure 3.2
Qualia as outputs of a comparator

label everything whose color looks similar to tree leaves. There will be lots of borderline cases, but that just reflects the boundaries in the underlying similarity space.

Now suppose we show the robot some pictures, and ask it to find the green object in each picture, if there is one. Obviously, this requires many visual skills besides the ability to label colors, but we are allowing ourselves to be ambitious for the purpose of these thought experiments. So we can assume the robot can find familiar object shapes in pictures, and can correctly assign colors to shapes.

Next, the robot looks at two pictures, and says that picture A contains no green objects and picture B contains a picture of a green house. Now we ask it to tell us how it knows that. It points to the house and says, "Look!" We explain that we are philosophers, and do not want the evidence; we want to know why the robot takes it to be evidence. The robot would say it doesn't know why. We point to a house in picture A, and ask, what's the difference between this and the green house in picture B? The robot would say, if forced to say anything, that house B "looks like" other green objects, and that house A "looks like" other blue objects. It could not possibly say anything else, because similarity judgments are the *only* link between the color-processing system and the language system. All green objects have something in common, the property of looking similar to each other, but as far as the robot is able to report, this is an irreducible property, a stopping point. The property they all have in common is right in front of the robot's face, but it can't be analyzed, at least not by the part of the robot that is connected to the speech system. *This something plays the role in the robot that the quale of the color plays in the human mind.*

The dust jacket of this book shows a painting by Bob Thompson (figure 3.3), an American artist from the 1960s who liked to paint scenes containing human forms in classical poses, but filled with monochromatic splashes of color instead of normal features. Colors would be repeated, so that two of the figures might be filled with the same bold yellow, while three others were filled with red. If a robot looked at this painting, it would see not just colors, but colors with a particular shape. If asked to comment on how it knew that shapes A and C were yellow whereas shapes B and D were red, it can only answer that shapes A and C are filled

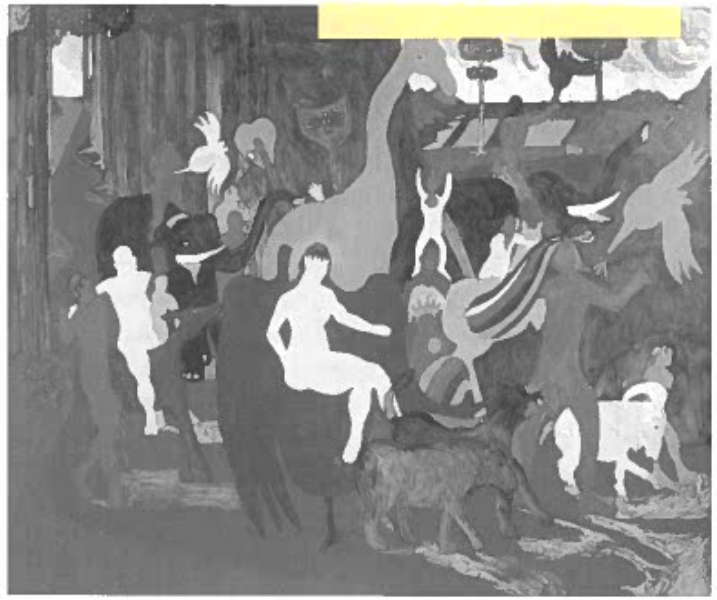

Figure 3.3
Bob Thompson, *Triumph of Bacchus,* 1964

with a something that looks one way, while B and D are filled with something that looks different. If asked to find other shapes in other paintings that resemble A or B, it would find shapes that were filled with the same "whatever it is" that fills A or B, or something "close."

My argument shows that if a robot were cognizant of judgments of color similarity that were structurally equivalent to humans', then they would be, as far as the robot could tell, introspectively equivalent, and hence the robot would believe that its experiences had qualia-like properties. However, as I have emphasized before, there is no reason that every robot would have to have judgments that were structurally equivalent to a human's. The robot might well have introspective access to color as a vector of three numbers (corresponding to the outputs of the three color filtors), or four numbers, or the Fourier transform of the light. The robot might classify colors in roughly the same way we do, but with significant differences in the similarity relationships of the type that have been discovered by psychophysicists (Clark 1993). But there is one point

in respect of which human and robot introspections must be alike: they must both draw a line somewhere between what is introspectable and what isn't. The robot's introspections about similarity judgments must eventually bottom out. If it represents numbers in binary notation, then it might experience colors as bit strings (labeling a spatial array of shapes, presumably). But then it would have no answer as to how it distinguishes a 1 from a 0. One would be experienced as ineffably "one-ish," and the other as exemplifying pure "zero-ness."

There are many ways in which a robot's introspections could differ structurally from humans'. A robot might be completely unable to say what color an object is in isolation. In other words, it might, when presented with a group of objects, be able to say how closely the pairs resemble each other, while refusing to grant that an object in isolation has anything like a color. It might be unable to discriminate a series of objects presented individually, while being able to discriminate them when presented simultaneously. People are not able to discriminate two similar shades of red unless they can see them side by side; now imagine a robot in a similar position toward red and green, or, for that matter, toward black and white. This is a possible design feature but doesn't seem likely to be included in a reasonable design, or to have evolved. In any case, even if such a robot's introspections were very different from humans', they would still be "experience-like." If the robot can become cognizant of what I called the "pictorial" qualities of its visual field, then it has to label the surfaces of the objects in the picture according to their intrinsic similarity relationships. If a label of something as intrinsically and unanalyzably similar to other things is not a quale, it is at least close to it.

Most of my examples have been drawn from the domain of vision, but the story is much the same in other areas. One peculiarity of many sensory systems is that their spatial field corresponds roughly to the entire body. A tickle is felt as being in a particular place on (or in) the body of the person being tickled. Another peculiarity is that some senses don't seem to convey any information beyond the fact that they're occurring. A tickle doesn't announce the presence of anything but a tickle. Vision, by contrast, normally simply delivers information about the physical positions of objects outside an organism, and it takes some effort to be cognizant

of the structure of the appearances of those objects separate from those objects.

Tye (1995) argues that this distinction is illusory and that the senses always represent something beyond their phenomenal quality. "What experiences of the tickle sort track (in optimal conditions) is the presence of something lightly touching or brushing against the surface of the body" (p. 116). I am willing to agree, with one qualification: it's really not necessary to be able to describe exactly what a sensory system tracks in order for it to have representational utility. Suppose a robot has three sorts of skin sensors, which react in different ways to different sorts of pressure. Having three different sensors might make it possible to perform most useful discriminations in most circumstances. That is, if there are two sorts of contact that are worth discriminating in some situations, the chances are good that not all three sensors will react to them in exactly the same way. Hence there is nothing in particular that the trio of sensors can be said to react to; what the sensors indicate in one environment may be quite different from what they indicate in a different environment. The discriminations would have to be learned, and might seldom rise to the level of reliability we associate with vision. In that case we would expect to see the boundary between perception and inference drawn differently, so that the perceiver is directly aware of the sensor readings and only inferentially aware of what is causing them. Indeed, that appears to be the case for touch as contrasted with vision, although there are counterexamples on both sides. A safecracker or heart surgeon no doubt becomes directly aware of events causing slight changes in the data received by touch. And, as Dennett (1991) has observed, our difficulties in saying exactly what physical property color *is* may derive from the fact that there is no need for the property to have any characterization other than "the property the human color vision system tracks." It suffices that many important differences in objects are correlated with differences in the way our color sensors react to them.

I said earlier that each emotion would involve a special quale in addition to a dimension of preference. Since we've established that robots assign qualia-like features to their cognizable percepts and assign different values to different outcomes, it seems inescapable that an intelligent robot would have emotion-like states. For example, if the expectation of danger and

the sensing of extreme heat are equally unpleasant but distinguishable to the robot, then the story the robot tells itself about how it does so involves ineffable qualities. So unpleasantness + quality 1 is a state we can label "fearlike"; while unpleasantess + quality 2 is "painlike" (for a particular kind of pain). These states play exactly the role emotions play in biological systems.

The Self-Model

We have examined a set of phenomena so far, choice, preference, and qualia, and in each case we can explain why robots would have to have them or something like them by appealing to the way complex computational systems would perceive their own processes of perception and action. In other words, these phenomena appear as features of a system's model of itself. The concept of self-model may appear somewhat mystical, as if I am conjuring consciousness by showing the reflection of a mirror. Let me hasten to demystify the idea completely.

A computational model C is a computational system that resembles a modeled system S in some respect and is used by a modeling system M to predict the behavior of S. A self-model arises when S = M. This may seem an unusual situation, but in fact it is common. Here are some examples:

- A computer that takes inventory of the furniture in an office may include itself in the inventory. If it is predicting future furniture needs, it may note that the computer (itself) will become obsolete and is planned to be replaced by a smaller, faster model in six months.
- A real-time compiler models the time required for various operations when it is producing code. The time depends on the computer the code will run on. Here S = M if the computer the code will run on is the same as the computer executing the compiler.
- A robot must decide how much planning to do before starting to carry its plan out. In some cases (Boddy and Dean 1989; Russell and Wefald 1991), it can use a statistical model of the expected benefit to be gained by further planning. It should spend only as much time planning as may be expected to yield an improved plan that saves more than that amount of time. Because the planner and the agent that will carry out the plan are the same, the statistical model qualifies as a self-model.

- Some robot hand-eye control systems look at their own grippers ("hands") as well as the objects they are manipulating. In figuring out the future trajectories of the objects, such a system must use a different model for objects it is gripping than for objects lying on the table. The self-model takes into account the movements the robot plans to undertake.

The last example is of a type that will become especially important as robots become more common. It attacks a problem that all animals face, namely, making sure they distinguish between self and nonself.

A key feature of humans' self-models is that they are unitary. Each of us models himself as a single person. Well, of course; that's what we are, aren't we? How could we model ourselves as anything else? Actually, as I have emphasized a couple of times, it is not at all clear that the way we think about ourselves is the only possible way, nor is it clear how many ways there are. Our brains consist of billions of neurons, and while it is implausible to imagine modeling them all, one could easily imagine modeling oneself as a community of modules (Minsky 1986). This is, at least in principle, independent of the question whether each of us *is* a community of modules. Models don't have to be perfectly accurate. A self-controlled spacecraft might model itself as a single rigid cylinder, even if its shape is really more complex, and even if it actually contains internal parts that are physically disconnected from the body of the spacecraft.

It is hard to convince the self of how unimportant it really is. How often have you had an experience like the following: One day I was headed for the men's bathroom on my floor, the fifth floor, but it was being cleaned. So I went to the bathroom one floor down. Now, it happens that one of the toilets on the fifth floor doesn't flush too well, and some member of the custodial staff has posted a sign over it, "For hygienic purposes, please flush!" (In vain, I might add.) I went to the fourth floor, walked into a bathroom that was almost identical to the one I usually go to, looked up, and was surprised to see that the sign was missing. Later, on my way out of the bathroom, I started to head for my office in its usual location, and had to change course and go back upstairs.

How shall we describe such a case? "I was absent-minded; I forgot where I was." True, but what was really going on, once I was in the bathroom, was that I was just behaving as I usually do. I didn't actually

believe I was still in the fifth-floor bathroom. If you had asked me where I was, I would have told you. My behavior was controlled by different subsystems than the ones that would have answered the question. From a conscious point of view this kind of event is inexplicable. Something other than "me" was in control. I think most people would be comfortable with that conclusion, but perhaps not the obvious corollary: that on occasions when I really am on the fifth floor, and my behavior is appropriate, my behavior is equally "inexplicable," although when we ask our self-model for an explanation it supplies one. If it is the self-model of a philosopher, it might say, "When I have a *desire* to go to the bathroom, and a *belief* that the bathroom is down the hall, I form the *intention* to go there, and then I go there." (See Milner 1999 for a remarkable list of cases where what people think they see and what they behave as if they see are quite different.)

Here is another sort of example. I have noticed that when I become skillful at a computer game (an activity induldged in for purely scientific reasons), the little creatures crawling across the screen seem to slow down. For instance, in the game of Gnibbles a worm crawls rapidly through a maze. If it hits a wall it dies. At first it seems impossible to control the worm. Before you can react, it crashes into something. However, eventually your nervous system gets "tuned" to the game. You anticipate what's going to happen, and now you seem to have all the time in the world to steer the worm left or right. But occasionally a situation pops up that you didn't anticipate, and before you can think the worm seems to speed up, spin out of control, and smash into the nearest obstacle. That's the way it seems, but the truth is that it the worm was never under conscious control, whatever that might mean. The self-model was just verifying that the worm was under control, and attributing that fact to decisions "you" made.

It is not hard to think of good reasons why we model ourselves as single persons in control of our minds and bodies. Each body can do just one thing at a time, or at least must carefully coordinate multiple activities. Multiple actions tend to occur as an ordered sequence. The brain module that controls needlework may have almost nothing in common with the one that controls tap dancing, but one must go first and the other second if the owner of these modules plans to do both.

Robots may not be under the same constraints as humans. It is not too farfetched to visualize teams of robots that act as a unit some of the time, and split into separate individuals the rest of the time. Their models of themselves might be very different from ours, although it is not necessary that they be; in principle, they could model themselves as a single creature with disconnected pieces.

Whatever the structure of a robot's self-model, the key point is that when it introspects it is "stuck" in that model. It can't escape from it. The model imposes certain basic boundary conditions on the questions it can ask. As I argued above, it can't stop believing that its actions are exempt from causal laws. It can't stop believing that certain things are intrinsically desirable or undesirable. It can't stop believing that objects are perceptually similar because of their intrinsic sensory qualities. Most important, it can't stop believing that it exists, or to make it more similar to what Descartes said, it can't stop believing "I exist." But what is meant by "I"? "I" is the creature who makes those free decisions, who feels attracted or repelled, or experiences the qualia of colors and sounds. In other words, "I" is an object in its self model, the key player as it were.[4]

It is a consequence of this theory that when the robot thinks about itself, it is manipulating a symbol that has meaning partly because of the model in which the manipulation takes place. Sometimes when a computational model is used to govern the behavior of a system, the symbols in the model end up denoting something because of the role they play in that behavior. Consider the file system on your personal computer; I mean the system of "documents," "folders," the "desktop," or similar entities that are used to organize data residing on your hard disk. If you were to print out the contents of the disk, you would get a long series of bits or characters. There would be some recognizable strings, but not in the correct order, and with lots of other junk interlarded. What makes this mess into a file system? It turns out that some of the junk is actually a description of how the pieces fit together. One block of bits describes how a bunch of other blocks (not necessarily contiguous) go together to make a file. Another block describes how a bunch of files are grouped into a folder. Other blocks specify where a folder will appear on the desktop, what aliases a file has, etc. A master block points to all the folders that do

not appear inside folders themselves, and when the computer boots one of the first things it does is to read this master block, whose location is fixed.

We don't normally think about this. We see an icon on the screen. It has a familiar design, so we know that, say, it is a word-processing document, and if we click on it its contents will appear in the arrangement the word processor produces. From our point of view, the icon denotes the file. From the programmers' point of view things are a bit different, and it's their point of view we're interested in. There is a data object representing the file (and another, which we don't care about right now, that denotes the icon). The denotation is as reliable as the computer repairperson can make it, but it has a peculiar feature: the existence of the data object is necessary to bring the denoted object into existence. Without such a data object the file would dissolve into a bunch of disconnected bits. In fact, files are usually deleted not by actually erasing anything but by just removing their descriptors from the appropriate data structures and declaring their blocks ready for use in new files.[5]

We see much the same pattern in the way the robot models the self. There is, we suppose, a symbol for "I." The robot would exhibit some behaviors whether it modeled itself or not, but some behaviors stem from the fact that it thinks about itself as a "person," that is, the fact that in its self-model the properties of a single entity with goals, emotions, and sensations are ascribed to "I." It behaves as such an entity because it models itself as such an entity; its behavior is to some extent constituted by the modeling.

The question is, does this "I" exist over and above the creature that has a model in which "I" exists? The answer is not quite as straightforward as the corresponding answer about the file system, but it is similar in form. The "I" in the model makes free decisions; does that bring a being with free will into existence? There are two rather different ways of answering the question, but for both the answer is yes. First, if the robot asks, does someone with free will really exist, the answer is, yes, I do! That's because the robot can't step out of the model. It may understand completely that it believes it is free only because it has a self-model with this belief, but that understanding does not allow it to escape the self-model and suspend that belief.

Second, the humans that interact with the robot, and other robots, for that matter, will perceive the owner of the model as being a creature with the same attributes it assigns to "I." People begin assuming at an early age that other people are making the kind of decisions and having the kind of experiences that they are having. A child learns the word for a concept by hearing the word when the thing it denotes is perceptually salient. Learning the word "choice" is no different from learning the word "chair." When the child is trying to decide between chocolate and vanilla, and her parents are urging her to make a "choice," then the cognitive state she is in gets labeled as a choice. Its classification as a type of state is prior to that point and presumably is an innate part of her self-model. When she later hears the word applied to other people, she assumes they are in a similar state of indecision. It is plausible to assume that intelligent robots, if they ever exist, will attribute to other robots the same properties they automatically perceive true of themselves.

Because of processes like this, intelligent robots would perceive other robots as selves similar to "I." In other words, one way the symbol "I" brings its denotation into being is by encouraging its owner to deal with other entities as though they all were creatures similar to the way it believes "I" to be. The robot's brain is making its "I" up as it goes along, but the process works the way it does partly because other intelligent systems are cooperating to make everyone else up too. Robot 1 believes Robot 2 to have (or be) a self like its own "I," so the self that the symbol "I" in Robot 2's model denotes is also denoted by whatever symbol Robot 1 uses for Robot 2.

There is one aspect of the self-model that we mustn't be too casual about. If we are not careful, the model will come to occupy the position of the spectator at the internal mental show in what Dennett (1991) calls the "Cartesian theater," the part that actually experiences. This is not the correct picture at all. There is no part that experiences. Experience inheres in the whole system, just as life inheres in a whole cell. A cell is alive but has no living parts, and the brain experiences but has no experiencing parts. The self-model is just another module in a collection of computational modules. It is fair to say that the self-model is a crucial component in the mechanism for maintaining the *illusion* that there is a Cartesian theater: it keeps track of the beliefs about the audience. To

carry out this role, it must have some special properties. One is that it is connected fairly directly to the parts of the system that are responsible for language. Another is that its conclusions are available for general-purpose inferences. This second property is stated somewhat vaguely, so much so that the first property might be a special case of it. The reason I am being vague is that I really don't know what "general-purpose inference" amounts to. But it seems as if a key purpose of introspection is the ability to acquire new capacities by reinterpreting sensory inputs. One learns that a stick that appears bent in water is really straight. The fact that it "appears bent" must be represented somewhere, and somehow associated with the inference that it "is actually straight." Where this association occurs is not known, but presumably it isn't the job of the module that normally measures the straightness of visible objects. It's not the job of an all-powerful self-model either. All the self-model does is reinterpret the input from other modules as information about perception and action, then feed it to where it can take part in inferences.

It may not be too early to speculate about where in the brain the self-model is located. Michel Gazzaniga (1998, p. 175) locates it in the left hemisphere, associated with the speech centers. He calls it "the interpreter."

> The insertion of an interpreter into an otherwise functioning brain delivers all kinds of by-products. A device that asks how infinite numbers of things relate to each other and gleans productive answers to that question can't help but give birth to the concept of self. Surely one of the questions that device would ask is "Who is solving these problems?" Call that "me," and away the problem goes!

However, it's likely that the self-model will not be a localized "black box" in the brain. The brain can't ship signals to arbitrary places to take part in computations the way electronic circuits do. If a computation involves two signals, the signals are usually generated close to where they will be used. Hence we would expect every piece of the brain to contain neurons that react to what the brain is doing as a brain activity instead of simply as a representation of events outside the body. Where the signals from these neurons go is a matter for speculation by someone who knows more about the brain than I do.

One key aspect of the self-model is that it seems to be connected to episodic memory. In our survey of AI in chapter 2, we touched on various

programs that learn things (such as maps and ways of winning games), but one thing they don't learn is what happened to them. TD-Gammon may play better backgammon because of a particular series of games it played, but it doesn't remember those games as particular events. The ability to remember and recall particular events is called *episodic memory*. People take this ability for granted, but it is really quite strange when you get down to it. Remembering an event is not simply recording a movie of it, not even a movie with synchronized sound track, smell track, and touch track. Memory is highly selective and not terribly reliable. One wonders why evolution would give rise to such a thing. Learning a skill, such as pitching horseshoes or playing backgammon, does not require remembering episodes in which that skill would have come in handy. Presumably simple organisms can only learn skills, and have no memory at all of the events along the way.

One plausible answer to the question of what episodic memory is for is that it supports learning when you don't know what you're learning. Something unusual happens and you remember the events just before it so that if the something unusual happens again you can see if the same kind of events preceded it the second time. "Remembering the events just before it" is vague and impossible to carry out completely, so you just store away the representations of a few of the events and hope for the best.

Episodic memory is not directly responsible for consciousness. But to the extent that the events in question are perceptual events, what will be remembered is the way things seemed. If you buy a new alarm clock, and are awakened the next morning by what sounds like your phone ringing, then you remember that the alarm clock sounded like the telephone. The next time you hear a similar sound (while lying in bed in the morning) you might remember that episode, and you might begin to clarify the differences between the sound of the telephone and the sound of the alarm.

One thing episodic memory and natural language have in common is that they seem to require general-purpose notations. If the brain doesn't know exactly why it's remembering something, it can't, as it were, "optimize" its notation for that purpose. It just has to strive to record "everything," even stuff that it might have thought was irrelevant. The natural-language system is similar in that it has to be able to talk about

"everything." I use quotes to remind you that the goal of a notation that can express everything is far-fetched and not well defined. It is highly unlikely that the brain comes close. Nonetheless, it comes closer here than anywhere else, and here is where the self-model plays a key role in giving it something to express.

The link between the self-model and natural language allows us to explain why "cognizance" is so closely tied to the ability to report, and allows us at the same time to break the link between them. We now see that to be cognizant of a state of affairs is for some representation of it to be accessible to the self-model. That is, one is cognizant of a state of affairs *A* if there is a representation of *A* such that that representation could itself be an object of perception (although on particular occasions it may never become one).[6] In the normal course of things, one can report about what is in the self-model, so it's not surprising that one can report about what one is cognizant of. However, in the case of a stroke or some other neurological problem the link can be broken, and cognizance could occur without the ability to report on anything. Some animals may have self-models even without language, and a robot certainly could be designed to have one without the other. If it seems hard to visualize, just reflect on the limits to one's own linguistic reporting ability. You can be cognizant of the difference between red and blue, but you can't describe it, except by pointing to red and blue things. Imagine having similar limits to other sorts of reports.

It is worthwhile to stop here and conduct a thought experiment about the strong connection between cognizance and language. What would a conscious entity be like if it could use language but not be able to talk about what it was aware of? At first glance it seems that there could be no such being. It is certainly hard to imagine what its mental life would be like. You might picture a sort of aphasia: such a creature would be able to talk about everything that was visible and tangible, but encounter some sort of block, or gap, when it tried to talk about its own experiences. It's hard to imagine this species as a real possibility, let alone as something that would be likely to evolve. The young would learn language perfectly well, rapidly assimilating words such as *cup* and *chair*. But when one of them tried to ask its parents questions about what it was experiencing, it would

draw a blank. It could not even utter the sentence, "Why can't I talk about some things, daddy?" Any sentence that even alluded to experience would be blocked. (I don't mean that the creature's tongue would feel physically prevented from speaking; I mean that, while the creature would have experiences, and be cognizant that it was having experiences, it would never feel tempted to talk about them.) It's hard to believe that a brain could filter out just this set of sentences, mainly because it's doubtful that the set is well defined.

There is, however, another possibility. Suppose there was a species on a faraway planet[7] for which *the entire linguistic apparatus* was disconnected from consciousness. These creatures can talk perfectly well, using languages just as rich as ours, but they do not know they can. Suppose creature A sees some buffalo and goes to tell creature B about it. He thinks (but not in words) *I must go see B*. B sees A coming. They stand near each other. After a while B realizes that A has seen some buffalo, and she decides what to do about it. However, B is unaware that A has *told* her about the buffalo. She knows that other people often bring information, but doesn't know how the information is transferred from one person to another. She knows that people often move their lips and make noises, but these motions and noises have social or sexual significance, and everyone is of the opinion that that's the only significance they have.

On this planet, scientists might eventually realize that the signals coming from mouths contain much more information than is generally assumed. By careful experimentation, they might figure out the code in which the information is carried. But figuring this out would not bring anything to consciousness. It would be analogous to our own investigations of neurons. We can, we believe, decipher the signals sent by neurons, but that doesn't make the content of those signals accessible to consciousness. Eventually the creatures' civilization might have a complete theory of linguistics, and understand perfectly how their language functions. But for them the whole system has as much to do with consciousness as digestion does for us.

An unconscious natural-language system like this is not hard to imagine. The computer programs that today carry on conversations on some topic are a far cry from full natural language, but one can imagine making

them much more sophisticated without making them conscious. What's hard to imagine is an unconscious language system existing *in addition* to consciousness. This species would have conscious thoughts that, by hypothesis, would not bear any resemblance to words. More precisely, no one would *know* whether they bore any resemblance to words. The issue would not come up until the creatures' science had advanced to the point where they knew that such things as words existed. Hence the contents of their consciousness would be like that of animals or small children, which we have such trouble visualizing.

Of course, it's very doubtful that a species like this could ever develop enough science to realize how their vocal apparatus transmitted information. If their language evolved in one medium, it would be extremely difficult to transfer it to a different medium; if their language was originally auditory, a written language would be very unlikely to develop. The creatures would think of communication as something that happened automatically and effortlessly, and only when two of them were standing near each other. They might imagine an object that could carry messages from one person to another, but they would imagine a magic rock, say, such that if one person stood next to it and thought something, then the next person who stood next to it would know what the first person had thought. Without a written language, it would be difficult to store and pass on the kind of detailed nonintuitive information that science consists of.

Could this race tell stories about magic rocks? In a way, yes. One of them might imagine a series of events involving such a rock. Then it might get passed on to the next person. It could be clearly marked as "hypothetical," so that the next person didn't think there actually *was* a magic rock. Soon everyone might be thinking about this hypothetical event sequence. The shared fantasy might contain a signal saying who the original author was, although in this culture it's doubtful that anyone would care, even assuming that the concept of "original author" had any meaning.

This species could not tell lies, however. Telling a lie involves an intention to get another creature to believe something that you know is false. Suppose a female of the species finds an attractive male and wants to conceal his location from another female. She might deliberately walk in the wrong direction or behave in some other misleading manner. However,

she couldn't tell the other female, "I think he's over there." Her unconscious communication system would make the decision what to say based on criteria that are inaccessible to her. The communication system might communicate false information, but this wouldn't count as lying. Suppose that natural selection has led to a situation in which information about the location of attractive mates is never transmitted. Then one day female A needs the assistance of female B in saving the life of male C, who is in some kind of mortal peril. A wants B to know the location of C, and runs to B. Alas, no matter how long she stands next to B, B will never know the location of C. A's brain might even send false information to B about C's location, but A isn't "lying" to B about where C is; A *wants* B to know where he is.

I'm sure everyone will agree that it is difficult to imagine the mental lives of creatures like this, but there is a worse, and deeper, problem. We take for granted that there is such a thing as "the self," so that when I sketch the situations above, and I say, "A thinks such-and-such," we automatically picture a self like ours having a thought like that. The problem is that our concepts of self, thought, and language are so intimately intertwined. If we move language from one realm to another, how do we know that the self stays put, rather than necessarily following the language facility? In the scenarios I discussed above, I assumed that the language facility would be inaccessible to the self, but can we be sure that the language facility knows that? Suppose the language of the creatures had words for "I" and "you," and these words were used consistently, even when talking about mental events. Suppose sentences such as these were produced:

- "You told me the apples were ripe, but most of them are still green."
- "From a distance, in that light, they looked red to me."
- "That story scared me."
- "Why did the magician cast the spell on the rock?"
- "Why did you tell me C was over there?"
- "I thought he was; I must have been mistaken."

One might rule them out, but it's not clear how. One might suppose that the sentence "They looked red to me" would be impossible, because language is disconnected from experience. However, if "why" questions cannot be asked and answered, it is not clear in what sense the creatures

have language at all. Besides, one can answer such questions without any "direct access" to consciousness, as shown by the fact that the questions work fine in the third person:

- "D told me the apples were ripe, but most of them are still green."
"From a distance, in that light, they looked red to D."
- "Why did D tell me that C was over there?"
"D thought he was; he must have been mistaken."

If an earth spaceship landed on this planet and found the inhabitants making sounds in each other's presence, the earthlings would naturally assume that the creatures were conversing. If they could decipher their language, they would hear conversations with sentences like those above. They would naturally assume that there was nothing odd about the creatures. They could talk to them as they would to any speaker of a new language. They would not realize that the "true selves" of this species were unaware that these conversations were taking place. Meanwhile, the "true selves" would find that the new creatures were at first unable to communicate, but after a while they would be able to absorb and transmit information just like the natives. The "true selves" would not, and presumably never could, realize that the earthlings' selves were connected to language the way they are.

Under these circumstances, it is not at all clear exactly how we should describe what's going on. Rather than say that the creatures' true selves are disconnected from language, perhaps it would be better to say that they each have two selves, one connected to language and the other not. If we can say that about them, then how do we know we can't say it about ourselves? How do we know there isn't a "true self" inside us that experiences lots of things, including some things we don't experience? The true self would know it got information from other people, but wouldn't realize that the transmission was mediated through vocal noises. There might even two or more inaccessible selves inside every person. Needless to say, these possibilities seem preposterous, but I think that ruling them out is impossible unless we have enough of a theory of consciousness to find the conscious systems in the world. This is a subject we will return to in chapters 4 and 5.

Virtual Consciousness and Real Consciousness

Throughout this chapter I have talked in terms of hypothetical intelligent robots and the way they would have to think of themselves if their thought processes were to be anything like ours. It's hard to guess what such robots would be like, assuming they could ever actually exist. Furthermore, the range of possibilities for the mental organization of a new genus of intelligent creatures is undoubtedly larger than we can imagine from the single data point that humans represent. However, I believe it is inescapable that robots would exhibit something *like* phenomenal consciousness. We can call it *virtual consciousness* to distinguish it, until proven otherwise, from the real thing. Virtual consciousness is the dependence on a self-model in which perceptions and emotions have qualia, some states of affairs are intrinsically better than others, and decisions are exempt from causal laws. Although there are currently no machines that exhibit virtual consciousness, the question of whether a machine or organism does exhibit it is purely a matter of third-person observation. It might be difficult to verify that it is present; the concept may require considerable revision as we understand intelligence better; but if our understanding advances as I expect, then testing whether a system exhibits *virtual* consciousness will eventually be completely uncontroversial, or at least only as controversial as testing whether a system has a belief.

What I would now like to claim is that real phenomenal consciousness and virtual phenomenal consciousness are indeed the same thing. Our brains maintain self-models with the required properties and that's why we think of ourselves, inescapably, as entities with emotions, sensations, and free will. When you have a sensation, you are representing a perceptual event using your self-model; when you make a decision, you are modeling yourself as exempt from causality; and so forth.

The evidence for this claim is simple, but it doesn't actually exist yet. I am anticipating the development of a more sophisticated cognitive science than we have now. When and if we have such a theory, I am assuming that it will involve many new computational constructs, but nothing above and beyond computation. What any given neuron does will be modelable as a computational process, in such a way that the neuron could be replaced by any other component that performed the same computation without

affecting the essential properties of the system. I am further assuming that self-models of the sort I am describing will be found in human brains, and probably the brains of other mammals.

If all this comes to pass, then we will be in a position to show without any doubt that virtual consciousness exists in human brains. There won't be any controversy about this, because virtual consciousness can be defined and investigated in purely "third-person" terms. Every *report* of a sensation, every *belief* in the freedom of a decision, will be accounted for in computational terms (and, by reduction, in neurological terms when the system under study is biological). The only way to deny that consciousness is identical with virtual consciousness will be to suppose that both are exhibited independently by the brain. Furthermore, in spite of our intuitions that when we report a sensation we are reporting on consciousness, it will be indubitable that the reports can actually be explained purely in terms of virtual consciousness. The belief that there is an additional process of consciousness will be very hard to sustain, especially given a demonstration that one aspect of virtual consciousness is the way it creates powerful, inescapable beliefs.

Of course, we are not in the position to make this argument yet, and we may never be. Many people may wish ardently that we never get there; I sometimes wish that myself. Nonetheless, if that's where we're going we might as well anticipate the consequences.

It is not easy to accept that the qualia-like entities robots believe in are in fact true qualia. When I experience the green of a tree, the key fact about it, besides its shape, is that that the shape is filled with "greenness." The robot is merely manipulating data structures. How in the world could those data structures exhibit a phenomenal quality like greenness? The key idea is that the robot has *beliefs* about the contents of its visual field, and the content of the beliefs is that a certain patch exists and is homogenously filled with something that marks it in some unanalyzable way as similar to other objects people call green. We do not, therefore, have to claim that the data structures themselves exhibit a quale, but simply that they support a belief in a quale. The relationship of quale to data structure is similar to the relationship between a fictional character and a book the character appears in. We don't expect a book about Godzilla to be taller than a building.

Please don't construe my proposal as a claim that "real consciousness doesn't exist; there's only the virtual kind." This would be analogous to concluding that nothing is really alive (because life is just a set of chemical reactions, no different in detail from chemical reactions in nonliving systems), or that nothing is really wet (because liquids, like solids, are really just atoms in motion, and atoms aren't wet). Consciousness is real, but turns out not to have all the properties we might have thought.

I have explained why a robot would model itself as having sensations, and why, in a sense, the model would be accurate. But I haven't quite said when it would be correct to say that a robot was having a sensation *now*. In other words, I need to explain "occurrent consciousess," as opposed to "the capacity for consciousness." The definition should be fairly obvious: A sensation is a particular perceptual event as modeled in the robot's self-model. Exactly what constitutes a particular perceptual event is not specified by the definition, but that's not important; it's whatever the model says it is. When the robot sees a sunset, it might in one instant be experiencing "sunset sensation," in the next a sensation caused by one cloud in the sunset, in the next the sensation of a particular spot of orange. Any particular occurrence of a modeled perceptual event is a sensation, just as any particular occurrence of a decision modeled as exempt from causal laws is an act of free will.

One consequence of this picture is that any perceptual event that is unmodeled does not involve a sensation. Obviously, perception does not cease simply because it is itself unperceived, but perception and consciousness are not the same thing. A thermostat reacts to high temperatures, but is not conscious of them. A thermostat that modeled itself in terms of sensations would, according to my theory, be conscious, but on occasions when a temperature measurement failed to make it to the self-model, there would be no more reason to suppose that it was conscious of that measurement than in the case of an ordinary thermostat.

For some people, all of this will be maddeningly beside the point, because it appears that I have simply neglected to explain what needs explaining, namely the actual qualitative character of my (or your) experiences. As Levine (1983, 1997) famously suggested, there is an "explanatory gap":

For a physicalist theory to be successful, it is not only necessary that it provide a physical description for mental states and properties, but also that it provide an *explanation* of these states and properties. In particular, we want an explanation of why when we occupy certain physico-functional states we experience qualitative character of the sort we do.... What is at issue is the ability to explain qualitative character itself; why it is like what it is like to see red or feel pain. (Levine 1997, p. 548)

Or it may appear that I have fallen into a simple confusion, mistaking what Block (1997*b*) calls "access consciousness" from the real target, phenomenal consciousness. A perception is access-conscious if it is "poised for direct control of thought and action." It is phenomenally conscious if it is *felt*, if it has a phenomenal character—a quale, in other words.

I stop at a traffic light. One of its bulbs means "stop," another means "go." Why does the "stop" light look like *this* and the "go" light look like *that*, instead of vice versa (or some other combination)? I use the demonstrative pronouns because the usual color words fail us here. You know what I mean: the two vivid qualia associated with stopping and going. How do *those particular qualities* follow from the computational theory of consciousness?

The answer is they don't; they couldn't. The theory explains why you have an ineradicable belief in those qualia, and *therefore* why there is nothing else to explain. When you think about your own mind, you use a self-model that supplies many beliefs about what's going on in that mind. The beliefs are generally useful, and generally close enough to the truth, but even when they are manufactured out of whole cloth they are still undoubtable, including the belief that "stop" lights look like *this* and "go" lights look like *that*.

Lycan (1997, p. 64) makes almost the same point this way:

My mental word [i.e., the symbol representing a sensation type] is functionally nothing like any of the complex expressions of English that in fact refer to the same (neural) state of affairs.... Since no one else can use that mental word ... to designate that state of affairs, of course no one can explain ... why that state of affairs feels like [that] to me.... Therefore, the lack of ... explanations, only to be expected, do not count against the materialist identification. They almost count in its favor.

You think you can imagine a world in which you experience different qualia for red and green objects, or in which my red quale is the same as your green. But what does it mean to compare qualia? If qualia exist only

in self-models, we have to explain what we mean by comparing entities in two disjoint self-models (those of two people or of one person in two possible worlds). But there is no such meaning to be had. The closest we can come is to imagine change within a single self-model, as when the colors switch and you can remember the way they used to be. (I will have more to say on this topic in chapter 4.)

My theory of phenomenal consciousness is in the tradition of what I called "second-order" theories in chapter 1. Such theories postulate that conscious thoughts are thoughts about or perceptions of "first-order" mental events that would otherwise be unconscious. In David Rosenthal's version (1986, 1997), the first-order mental events are nonconscious thoughts, and the second-order events are thoughts about them. However, he would resist the identification of "thoughts" with computational entities. Lycan (1987) (see also Lycan 1996, 1997), following Armstrong (1968), proposes that consciousness is a matter of "self-scanning," or "inner perception." But Lycan (1997, p. 76) believes this idea explains only subjectivity (perhaps the same as Block's "access consciousness"), and not qualia: "... The mere addition of a higher-order monitoring to an entirely nonqualitative mental state could not possibly bring a *quale* into being.... The monitoring only makes the subject aware of a quale that was there, independently, in the first place" (Lycan 1996, pp. 76–77). This I emphatically deny. There is simply no place, and no need, for qualia in an ordinary computational system. The quale is brought into being solely by the process of self-modeling.

Georges Rey (1997) states the key insight reluctantly but convincingly thus:

We might ... include [in an intelligent machine] ... sensors that would signal to the machine the presence of certain kinds of damage to its surface or parts of its interior. These signals could be processed in such a way as to cause in the machine a sudden, extremely high preference assignment, to the implementation of any sub-routine that the machine believed likely to reduce that damage and/or the further reception of such signals.... The states produced in this way would seem to constitute the functional equivalent of pain.... Most of us would pretty surely balk at claiming that [such] a machine ... should be regarded as *really* having the experience of red just because it has a transducer that emits a characteristic signal, with some of the usual cognitive consequences, whenever it is stimulated with red light. But I'm not sure what entitles us to our reservations. For what else is there? In particular, what else is there that we are so sure is there and essential

> in our own case? . . . How do we know *we* "have the experience of red" over and above our undergoing just such a process as I have described in this machine?
>
> The machine could think and print out "I see clearly that there is nothing easier for me to know than my own mind," and proceed to insist that "no matter what your theory and instruments might say, they can never give me reason to think that I am not conscious here, now." . . . If someone now replies that we've only provided the machine with the functional equivalent of consciousness, we may ask. . . , what more is required?[8] (pp. 470–471)

The idea that consciousness arises through the use of a self-model has also been put forth by Minsky (1968), Hofstadter and Dennett (1981), Dennett (1991), and Dawkins (1989).

There is a longer list of people who disagree with this whole family of theories. In the next chapter I will discuss and refute their objections.