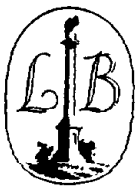
CONSCIOUSNESS EXPLAINED

DANIEL C. DENNETT

Illustrated by Paul Weiner



LITTLE, BROWN AND COMPANY

BOSTON TORONTO LONDON

1

PRELUDE: HOW ARE HALLUCINATIONS POSSIBLE?

1. THE BRAIN IN THE VAT

Suppose evil scientists removed your brain from your body while you slept, and set it up in a life-support system in a vat. Suppose they then set out to trick you into believing that you were not just a brain in a vat, but still up and about, engaging in a normally embodied round of activities in the real world. This old saw, the brain in the vat, is a favorite thought experiment in the toolkit of many philosophers. It is a modern-day version of Descartes's (1641)[1] evil demon, an imagined illusionist bent on tricking Descartes about absolutely everything, including his own existence. But as Descartes observed, even an infinitely powerful evil demon couldn't trick him into thinking he himself existed if he didn't exist: cogito ergo sum, "I think, therefore I am." Philosophers today are less concerned with proving one's own existence as a thinking thing (perhaps because they have decided that Descartes settled that matter quite satisfactorily) and more concerned about what, in principle, we may conclude from our experience about our nature, and about the nature of the world in which we (apparently) live. *Might* you be nothing but a brain in a vat? Might you have *always* been just a brain in a vat? If so, could you even conceive of your predicament (let alone confirm it)?

The idea of the brain in the vat is a vivid way of exploring these questions, but I want to put the old saw to another use. I want to use

1. Dates in parentheses refer to works listed in the Bibliography.

it to uncover some curious facts about hallucinations, which in turn will lead us to the beginnings of a theory — an empirical, scientifically respectable theory — of human consciousness. In the standard thought experiment, it is obvious that the scientists would have their hands full providing the nerve stumps from all your senses with just the right stimulations to carry off the trickery, but philosophers have assumed for the sake of argument that however technically difficult the task might be, it is "possible in principle." One should be leery of these possibilities in principle. It is also possible in principle to build a stainless-steel ladder to the moon, and to write out, in alphabetical order, all intelligible English conversations consisting of less than a thousand words. But neither of these are remotely possible in fact and sometimes an *impossibility in fact* is theoretically more interesting than a *possibility in principle*, as we shall see.

Let's take a moment to consider, then, just how daunting the task facing the evil scientists would be. We can imagine them building up to the hard tasks from some easy beginnings. They begin with a conveniently comatose brain, kept alive but lacking all input from the optic nerves, the auditory nerves, the somatosensory nerves, and all the other afferent, or input, paths to the brain. It is sometimes assumed that such a "deafferented" brain would naturally stay in a comatose state forever, needing no morphine to keep it dormant, but there is some empirical evidence to suggest that spontaneous waking might still occur in these dire circumstances. I think we can suppose that were you to awake in such a state, you would find yourself in horrible straits: blind, deaf, completely numb, with no sense of your body's orientation.

Not wanting to horrify you, then, the scientists arrange to wake you up by piping stereo music (suitably encoded as nerve impulses) into your auditory nerves. They also arrange for the signals that would normally come from your vestibular system or inner ear to indicate that you are lying on your back, but otherwise paralyzed, numb, blind. This much should be within the limits of technical virtuosity in the near future — perhaps possible even today. They might then go on to stimulate the tracts that used to innervate your epidermis, providing it with the input that would normally have been produced by a gentle, even warmth over the ventral (belly) surface of your body, and (getting fancier) they might stimulate the dorsal (back) epidermal nerves in a way that simulated the tingly texture of grains of sand pressing into your back. "Great!" you say to yourself: "Here I am, lying on my back on

the beach, paralyzed and blind, listening to rather nice music, but probably in danger of sunburn. How did I get here, and how can I call for help?"

But now suppose the scientists, having accomplished all this, tackle the more difficult problem of convincing you that you are not a mere beach potato, but an agent capable of engaging in some form of activity in the world. Starting with little steps, they decide to lift part of the "paralysis" of your phantom body and let you wiggle your right index finger in the sand. They permit the sensory experience of moving your finger to occur, which is accomplished by giving you the kinesthetic feedback associated with the relevant volitional or motor signals in the output or efferent part of your nervous system, but they must also arrange to remove the numbness from your phantom finger, and provide the stimulation for the feeling that the motion of the imaginary sand around your finger would provoke.

Suddenly, they are faced with a problem that will quickly get out of hand, for just how the sand will feel depends on just how you decide to move your finger. The problem of calculating the proper feedback, generating or composing it, and then presenting it to you in real time is going to be computationally intractable on even the fastest computer, and if the evil scientists decide to solve the real-time problem by precalculating and "canning" all the possible responses for playback, they will just trade one insoluble problem for another: there are too many possibilities to store. In short, our evil scientists will be swamped by *combinatorial explosion* as soon as they give you any genuine exploratory powers in this imaginary world.[2]

It is a familiar wall these scientists have hit; we see its shadow in the boring stereotypes in every video game. The alternatives open

---

2. The term *combinatorial explosion* comes from computer science, but the phenomenon was recognized long before computers, for instance in the fable of the emperor who agrees to reward the peasant who saved his life one grain of rice on the first square of the checkerboard, two grains on the second, four on the third, and so forth, doubling the amount for each of the sixty-four squares. He ends up owing the wily peasant millions of billions of grains of rice (2^{64} to be exact). Closer to our example is the plight of the French "aleatoric" novelists who set out to write novels in which, after reading chapter 1, the reader flips a coin and then reads chapter 2a or 2b, depending on the outcome, and then reads chapter 3aa, 3ab, 3ba, or 3bb after that, and so on, flipping a coin at the end of every chapter. These novelists soon came to realize that they had better minimize the number of choice points if they wanted to avoid an explosion of fiction that would prevent anyone from carrying the whole "book" home from the bookstore.

for action have to be strictly — and unrealistically — limited to keep the task of the world-representers within feasible bounds. If the scientists can do no better than convince you that you are doomed to a lifetime of playing Donkey Kong, they are evil scientists indeed.

There is a solution of sorts to this technical problem. It is the solution used, for instance, to ease the computational burden in highly realistic flight simulators: use *replicas* of the items in the simulated world. Use a real cockpit and push and pull it with hydraulic lifters, instead of trying to simulate all that input to the seat of the pants of the pilot in training. In short, there is only one way for you to store for ready access that much information about an imaginary world to be explored, and that is to use a *real* (if tiny or artificial or plaster-of-paris) world to store its own information! This is "cheating" if you're the evil demon claiming to have deceived Descartes about the existence of absolutely everything, but it's a way of actually getting the job done with less than infinite resources.

Descartes was wise to endow his imagined evil demon with *infinite* powers of trickery. Although the task is not, strictly speaking, infinite, the amount of information obtainable in short order by an inquisitive human being is staggeringly large. Engineers measure information flow in bits per second, or speak of the *bandwidth* of the channels through which the information flows. Television requires a greater bandwidth than radio, and high-definition television has a still greater bandwidth. High-definition smello-feelo television would have a still greater bandwidth, and *interactive* smello-feelo television would have an astronomical bandwidth, because it constantly branches into thousands of slightly different trajectories through the (imaginary) world. Throw a skeptic a dubious coin, and in a second or two of hefting, scratching, ringing, tasting, and just plain looking at how the sun glints on its surface, the skeptic will consume more bits of information than a Cray supercomputer can organize in a year. Making a *real* but counterfeit coin is child's play; making a *simulated* coin out of nothing but organized nerve stimulations is beyond human technology now and probably forever.[3]

---

3. The development of "Virtual Reality" systems for recreation and research is currently undergoing a boom. The state of the art is impressive: electronically rigged gloves that provide a convincing interface for "manipulating" virtual objects, and head-mounted visual displays that permit you to explore virtual environments of considerable complexity. The limitations of these systems are apparent, however, and they bear out

One conclusion we can draw from this is that we are not brains in vats — in case you were worried. Another conclusion it seems that we can draw from this is that strong hallucinations are simply impossible! By a strong hallucination I mean a hallucination of an apparently concrete and persisting three-dimensional object in the real world — as contrasted to flashes, geometric distortions, auras, afterimages, fleeting phantom-limb experiences, and other anomalous sensations. A strong hallucination would be, say, a ghost that talked back, that permitted you to touch it, that resisted with a sense of solidity, that cast a shadow, that was visible from any angle so that you might walk around it and see what its back looked like.

Hallucinations can be roughly ranked in strength by the number of such features they have. Reports of very strong hallucinations are rare, and we can now see why it is no coincidence that the credibility of such reports seems, intuitively, to be inversely proportional to the strength of the hallucination reported. We are — and should be — particularly skeptical of reports of very strong hallucinations because we don't believe in ghosts, and we think that only a real ghost could produce a strong hallucination. (It was primarily the telltale strength of the hallucinations reported by Carlos Castañeda in *The Teachings of Don Juan: A Yaqui Way of Knowledge* [1968] that first suggested to scientists that the book, in spite of having been a successful Ph.D. thesis in anthropology at UCLA, was fiction, not fact.)

But if *really* strong hallucinations are not known to occur, there can be no doubt that convincing, multimodal hallucinations are frequently experienced. The hallucinations that are well attested in the literature of clinical psychology are often detailed fantasies far beyond the generative capacities of current technology. How on earth can a single brain do what teams of scientists and computer animators would find to be almost impossible? If such experiences are not genuine or veridical perceptions of some real thing "outside" the mind, they must be produced entirely inside the mind (or the brain), concocted out of whole cloth but lifelike enough to fool the very mind that concocts them.

---

my point: it is only by various combinations of physical replicas and schematization (a *relatively* coarse-grained representation) that robust illusions can be sustained. And even at their best, they are experiences of virtual surreality, not something that you might mistake for the real thing for more than a moment. If you really want to fool someone into thinking he is in a cage with a gorilla, enlisting the help of an actor in a gorilla suit is going to be your best bet for a long time.

2. PRANKSTERS IN THE BRAIN

The standard way of thinking of this is to suppose that hallucinations occur when there is some sort of freakish autostimulation of the brain, in particular, an entirely internally generated stimulation of some parts or levels of the brain's perceptual systems. Descartes, in the seventeenth century, saw this prospect quite clearly, in his discussion of phantom limb, the startling but quite normal hallucination in which amputees seem to feel not just the presence of the amputated part, but itches and tingles and pains in it. (It often happens that new amputees, after surgery, simply cannot believe that a leg or foot has been amputated until they see that it is gone, so vivid and realistic are their sensations of its continued presence.) Descartes's analogy was the bell-pull. Before there were electric bells, intercoms, and walkie-talkies, great houses were equipped with marvelous systems of wires and pulleys that permitted one to call for a servant from any room in the house. A sharp tug on the velvet sash dangling from a hole in the wall pulled a wire that ran over pulleys all the way to the pantry, where it jangled one of a number of labeled bells, informing the butler that service was required in the master bedroom or the parlor or the billiards room. The systems worked well, but were tailor-made for pranks. Tugging on the parlor wire anywhere along its length would send the butler scurrying to the parlor, under the heartfelt misapprehension that someone had called him from there — a modest little hallucination of sorts. Similarly, Descartes thought, since perceptions are caused by various complicated chains of events in the nervous system that lead eventually to the control center of the conscious mind, if one could intervene somewhere along the chain (anywhere on the optic nerve, for instance, between the eyeball and consciousness), tugging just right on the nerves would produce exactly the chain of events that would be caused by a normal, veridical perception of something, and this would produce, at the receiving end in the mind, exactly the effect of such a conscious perception.

The brain — or some part of it — inadvertently played a mechanical trick on the mind. That was Descartes's explanation of phantom-limb hallucinations. Phantom-limb hallucinations, while remarkably vivid, are — by our terminology — relatively weak; they consist of unorganized pains and itches, all in one sensory modality. Amputees don't see or hear or (so far as I know) smell their phantom feet. So something like Descartes's account could be the right way to explain phantom limbs, setting aside for the time being the notorious mysteries about

how the physical brain could interact with the nonphysical conscious mind. But we can see that even the purely mechanical part of Descartes's story must be wrong as an account of relatively strong hallucinations; there is no way the brain as illusionist could store and manipulate enough false information to fool an inquiring mind. The brain can relax, and let the real world provide a surfeit of true information, but if it starts trying to short-circuit its own nerves (or pull its own wires, as Descartes would have said), the results will be only the weakest of fleeting hallucinations. (Similarly, the malfunctioning of your neighbor's electric hairdryer might cause "snow" or "static," or hums and buzzes, or odd flashes to appear on your television set, but if you see a bogus version of the evening news, you *know* it had an elaborately organized cause far beyond the talents of a hairdryer.)

It is tempting to suppose that perhaps we have been too gullible about hallucinations; perhaps only mild, fleeting, thin hallucinations ever occur — the strong ones don't occur because they can't occur! A cursory review of the literature on hallucinations certainly does suggest that there is something of an inverse relation between strength and frequency — as well as between strength and credibility. But that review also provides a clue leading to another theory of the mechanism of hallucination-production: one of the endemic features of hallucination reports is that the victim will comment on his or her rather unusual passivity in the face of the hallucination. Hallucinators usually just stand and marvel. Typically, they feel no desire to probe, challenge, or query, and take no steps to interact with the apparitions. It is likely, for the reasons we have just explored, that this passivity is not an inessential feature of hallucination but a necessary precondition for any moderately detailed and sustained hallucination to occur.

Passivity, however, is only a special case of a way in which relatively strong hallucinations could survive. The reason these hallucinations can survive is that the illusionist — meaning by that, whatever it is that produces the hallucination — can "count on" a particular line of exploration by the victim — in the case of total passivity, the *null* line of exploration. So long as the illusionist can predict in detail the line of exploration actually to be taken, it only has to prepare for the illusion to be sustained "in the directions that the victim will look." Cinema set designers insist on knowing the location of the camera in advance — or if it is not going to be stationary, its exact trajectory and angle — for then they have to prepare only enough material to cover the perspectives actually taken. (Not for nothing does *cinéma verité* make extensive use of the freely roaming hand-held camera.) In real

life the same principle was used by Potemkin to economize on the show villages to be reviewed by Catherine the Great; her itinerary had to be ironclad.

So one solution to the problem of strong hallucination is to suppose that there is a link between the victim and illusionist that makes it possible for the illusionist to build the illusion *dependent on*, and hence capable of anticipating, the exploratory intentions and decisions of the victim. Where the illusionist is unable to "read the victim's mind" in order to obtain this information, it is still sometimes possible in real life for an illusionist (a stage magician, for instance) to *entrain* a particular line of inquiry through subtle but powerful "psychological forcing." Thus a card magician has many standard ways of giving the victim the illusion that he is exercising his free choice in what cards on the table he examines, when in fact there is only one card that may be turned over. To revert to our earlier thought experiment, if the evil scientists can *force* the brain in the vat to have a particular set of exploratory intentions, they can solve the combinatorial explosion problem by preparing only the anticipated material; the system will be only *apparently* interactive. Similarly, Descartes's evil demon can sustain the illusion with less than infinite power if he can sustain an illusion of free will in the victim, whose investigation of the imaginary world he minutely controls.[4]

But there is an even more economical (and realistic) way in which hallucinations could be produced in a brain, a way that harnesses the very freewheeling curiosity of the victim. We can understand how it works by analogy with a party game.

3. A PARTY GAME CALLED PSYCHOANALYSIS

In this game one person, the dupe, is told that while he is out of the room, one member of the assembled party will be called upon to relate a recent dream. This will give everybody else in the room the story line of that dream so that when the dupe returns to the room and begins questioning the assembled party, the dreamer's identity will be hidden in the crowd of responders. The dupe's job is to ask yes/no questions of the assembled group until he has figured out the dream narrative to a suitable degree of detail, at which point the dupe is to

4. For a more detailed discussion of the issues of free will, control, mindreading, and anticipation, see my *Elbow Room: The Varieties of Free Will Worth Wanting*, 1984, especially chapters 3 and 4.

psychoanalyze the dreamer, and use the analysis to identify him or her.

Once the dupe is out of the room, the host explains to the rest of the party that no one is to relate a dream, that the party is to answer the dupe's questions according to the following simple rule: if the last letter of the last word of the question is in the first half of the alphabet, the questions is to be answered in the affirmative, and all other questions are to be answered in the negative, with one proviso: a noncontradiction override rule to the effect that later questions are not to be given answers that contradict earlier answers. For example:

Q: Is the dream about a girl?
A: Yes.

but if later our forgetful dupe asks

Q: Are there any female characters in it?
A: Yes [in spite of the final t, applying the noncontradiction override rule].[5]

When the dupe returns to the room and begins questioning, he gets a more or less random, or at any rate arbitrary, series of yeses and noes in response. The results are often entertaining. Sometimes the process terminates swiftly in absurdity, as one can see at a glance by supposing the initial question asked were "Is the story line of the dream word-for-word identical to the story line of *War and Peace*?" or, alternatively, "Are there any animate beings in it?" A more usual outcome is for a bizarre and often obscene story of ludicrous misadventure to unfold, to the amusement of all. When the dupe eventually decides that the dreamer — whoever he or she is — must be a very sick and troubled individual, the assembled party gleefully retorts that the dupe himself is the author of the "dream." This is not strictly true, of course. In one sense, the dupe is the author by virtue of the questions he was inspired to ask. (No one else proposed putting the three gorillas in the rowboat with the nun.) But in another sense, the dream simply has no author, and that is the whole point. Here we see a process of narrative production, of detail accumulation, with no authorial intentions or plans at all — an illusion with no illusionist.

The structure of this party game bears a striking resemblance to the structure of a family of well-regarded models of perceptual systems.

---

5. Empirical testing suggests that the game is more likely to produce a good story if in fact you favor affirmative answers slightly, by making p/q the alphabetic dividing line between yes and no.

It is widely held that human vision, for instance, cannot be explained as an *entirely* "data-driven" or "bottom-up" process, but needs, at the highest levels, to be supplemented by a few "expectation-driven" rounds of hypothesis testing (or something analogous to hypothesis testing). Another member of the family is the "analysis-by-synthesis" model of perception that also supposes that perceptions are built up in a process that weaves back and forth between centrally generated expectations, on the one hand, and confirmations (and disconfirmations) arising from the periphery on the other hand (e.g., Neisser, 1967). The general idea of these theories is that after a certain amount of "preprocessing" has occurred in the early or peripheral layers of the perceptual system, the tasks of perception are completed — objects are identified, recognized, categorized — by generate-and-test cycles. In such a cycle, one's current expectations and interests shape hypotheses for one's perceptual systems to confirm or disconfirm, and a rapid sequence of such hypothesis generations and confirmations produces the ultimate product, the ongoing, updated "model" of the world of the perceiver. Such accounts of perception are motivated by a variety of considerations, both biological and epistemological, and while I wouldn't say that any such model has been proven, experiments inspired by the approach have borne up well. Some theorists have been so bold as to claim that perception must have this fundamental structure.

Whatever the ultimate verdict turns out to be on generate-and-test theories of perception, we can see that they support a simple and powerful account of hallucination. All we need suppose must happen for an otherwise normal perceptual system to be thrown into a hallucinatory mode is for the hypothesis-generation side of the cycle (the expectation-driven side) to operate normally, while the data-driven side of the cycle (the confirmation side) goes into a disordered or random or arbitrary round of confirmation and disconfirmation, just as in the party game. In other words, if noise in the data channel is arbitrarily amplified into "confirmations" and "disconfirmations" (the arbitrary yes and no answers in the party game), the current expectations, concerns, obsessions, and worries of the victim will lead to framing questions or hypotheses whose content is guaranteed to reflect those interests, and so a "story" will unfold in the perceptual system without an author. We don't have to suppose the story is written in advance; we don't have to suppose that information is stored or composed in the illusionist part of the brain. All we suppose is that the illusionist

goes into an arbitrary confirmation mode and the victim provides the content by asking the questions.

This provides in the most direct possible way a link between the emotional state of the hallucinator and the content of the hallucinations produced. Hallucinations are usually related in their content to the current concerns of the hallucinator, and this model of hallucination provides for that feature without the intervention of an implausibly knowledgeable internal storyteller who has a theory or model of the victim's psychology. Why, for instance, does the hunter on the last day of deer season see a deer, complete with antlers and white tail, while looking at a black cow or another hunter in an orange jacket? Because his internal questioner is obsessively asking: "Is it a deer?" and getting NO for an answer until finally a bit of noise in the system gets mistakenly amplified into a YES, with catastrophic results.

A number of findings fit nicely with this picture of hallucination. For instance, it is well known that hallucinations are the normal result of prolonged sensory deprivation (see, e.g., Vosberg, Fraser, and Guehl, 1960). A plausible explanation of this is that in sensory deprivation, the data-driven side of the hypothesis-generation-and-test system, lacking any data, lowers its threshold for noise, which then gets amplified into arbitrary patterns of confirmation and disconfirmation signals, producing, eventually, detailed hallucinations whose content is the product of nothing more than anxious expectation and chance confirmation. Moreover, in most reports, hallucinations are only gradually elaborated (under conditions of either sensory deprivation or drugs). They start out weak — e.g., geometric — and then become stronger ("objective" or "narrative"), and this is just what this model would predict (see, e.g., Siegel and West, 1975).

Finally, the mere fact that a drug, by diffusion in the nervous system, can produce such elaborate and contentful effects requires explanation — the drug *itself* surely can't "contain the story," even if some credulous people like to think so. It is implausible that a drug, by diffuse activity, could create or even turn on an elaborate illusionist system, while it is easy to see how a drug could act directly to raise or lower or disorder in some arbitrary way a confirmation threshold in a hypothesis-generation system.

The model of hallucination generation inspired by the party game could also explain the composition of dreams, of course. Ever since Freud there has been little doubt that the thematic content of dreams is tellingly symptomatic of the deepest drives, anxieties, and

preoccupations of the dreamer, but the clues the dreams provide are notoriously well concealed under layers of symbolism and misdirection. What kind of process could produce stories that speak so effectively and incessantly to a dreamer's deepest concerns, while clothing the whole business in layers of metaphor and displacement? The more or less standard answer of the Freudian has been the extravagant hypothesis of an internal dream playwright composing therapeutic dramaplays for the benefit of the ego and cunningly sneaking them past an internal censor by disguising their true meaning. (We might call the Freudian model the Hamlet model, for it is reminiscent of Hamlet's devious ploy of staging "The Mousetrap" just for Claudius; it takes a clever devil indeed to dream up such a subtle stratagem, but if Freud is to be believed, we all harbor such narrative virtuosi.) As we shall see later on, theories that posit such *homunculi* ("little men" in the brain) are not always to be shunned, but whenever homunculi are rung in to help, they had better be relatively stupid functionaries — not like the brilliant Freudian playwrights who are supposed to produce new dream-scenes every night for each of us! The model we are considering eliminates the playwright altogether, and counts on the "audience" (analogous to the one who is "it" in the party game) to provide the content. The audience is no dummy, of course, but at least it doesn't have to have a theory of its own anxieties; it just has to be driven by them to ask questions.

It is interesting to note, by the way, that one feature of the party game that would not be necessary for a process producing dreams or hallucinations is the noncontradiction override rule. Since one's perceptual systems are presumably always exploring an ongoing situation (rather than a *fait accompli*, a finished dream narrative already told) subsequent "contradictory" confirmations can be interpreted by the machinery as indicating a new change in the world, rather than a revision in the story known by the dream relaters. The ghost was blue when last I looked, but has now suddenly turned green; its hands have turned into claws, and so forth. The volatility of metamorphosis of objects in dreams and hallucinations is one of the most striking features of those narratives, and what is even more striking is how seldom these noticed metamorphoses "bother" us while we are dreaming. So the farmhouse in Vermont is now suddenly revealed to be a bank in Puerto Rico, and the horse I was riding is now a car, no a speedboat, and my companion began the ride as my grandmother but has become the Pope. These things happen.

This volatility is just what we would expect from an active but

sufficiently skeptical question-asker confronted by a random sample f yeses and noes. At the same time, the persistence of some themes nd objects in dreams, their refusal to metamorphose or disappear, can lso be tidily explained by our model. Pretending, for the moment, that he brain uses the alphabet rule and conducts its processing in English, ve can imagine how subterranean questioning goes to create an obses- ive dream:

Q. Is it about father?
A. No.
Q. Is it about a telephone?
A. Yes.
Q. Okay. Is it about mother?
A. No.
Q. Is it about father?
A. No.
Q. Is it about father on the telephone?
A. Yes.
Q. I knew it was about father! Now, was he talking to me?
A. Yes. . . .

This little theory sketch could hardly be said to prove anything (yet) about hallucinations or dreams. It does show — metaphorically — how a mechanistic explanation of these phenomena might go, and that's an important prelude, since some people are tempted by the defeatist thesis that science couldn't "in principle" explain the various "mysteries" of the mind. The sketch so far, however, does not even address the problem of our consciousness of dreams and hallucinations. Moreover, although we have exorcised one unlikely homunculus, the clever illusionist/playwright who plays pranks on the mind, we have left in his place not only the stupid question-answerers (who arguably can be "replaced by machines") but also the still quite clever and unexplained question-poser, the "audience." If we have eliminated a villain, we haven't even begun to give an account of the victim.

We have made some progress, however. We have seen how atten- tion to the "engineering" requirements of a mental phenomenon can raise new, and more readily answerable, questions, such as: What models of hallucination can avoid combinatorial explosion? How might the content of experience be elaborated by (relatively) stupid, uncom- prehending processes? What sort of links between processes or systems could explain the results of their interaction? If we are to compose a

scientific theory of consciousness, we will have to address many questions of this sort.

We have also introduced a central idea in what is to follow. The key element in our various explanations of how hallucinations and dreams are possible at all was the theme that the only work that the brain must do is whatever it takes to assuage epistemic hunger — to satisfy "curiosity" in all its forms. If the "victim" is passive or incurious about topic x, if the victim doesn't seek answers to any questions about topic x, then no material about topic x needs to be prepared. (Where it doesn't itch, don't scratch.) The world provides an inexhaustible deluge of information bombarding our senses, and when we concentrate on how much is coming in, or continuously available, we often succumb to the illusion that it all must be used, all the time. But our capacities to use information, and our epistemic appetites, are limited. If our brains can just satisfy all our particular epistemic hungers as they arise, we will never find grounds for complaint. We will never be able to tell, in fact, that our brains are provisioning us with less than everything that is available in the world.

So far, this thrifty principle has only been introduced, not established. As we shall see, the brain doesn't always avail itself of this option in any case, but it's important not to overlook the possibility. The power of this principle to dissolve ancient conundrums has not been generally recognized.

4. PREVIEW

In the chapters that follow, I will attempt to explain consciousness. More precisely, I will explain the various phenomena that compose what we call consciousness, showing how they are all physical effects of the brain's activities, how these activities evolved, and how they give rise to illusions about their own powers and properties. It is very hard to imagine how your mind could be your brain — but not impossible. In order to imagine this, you really have to know quite a lot of what science has discovered about how brains work, but much more important, you have to learn new ways of thinking. Adding facts helps you imagine new possibilities, but the discoveries and theories of neuroscience are not enough — even neuroscientists are often baffled by consciousness. In order to stretch your imagination, I will provide, along with the relevant scientific facts, a series of stories, analogies, thought experiments, and other devices designed to give you new per-

spectives, break old habits of thought, and help you organize the facts into a single, coherent vision strikingly different from the traditional view of consciousness we tend to trust. The thought experiment about the brain in the vat and the analogy with the game of psychoanalysis are warm-up exercises for the main task, which is to sketch a theory of the biological mechanisms *and a way of thinking* about these mechanisms that will let you see how the traditional paradoxes and mysteries of consciousness can be resolved.

In Part I, we survey the problems of consciousness and establish some methods. This is more important and difficult than one might think. Many of the problems encountered by other theories are the result of getting off on the wrong foot, trying to guess the answers to the Big Questions too early. The novel background assumptions of my theory play a large role in what follows, permitting us to *postpone* many of the traditional philosophical puzzles over which other theorists stumble, until after we have outlined an empirically based theory, which is presented in Part II.

The Multiple Drafts model of consciousness outlined in Part II is an alternative to the traditional model, which I call the Cartesian Theater. It requires a quite radical rethinking of the familiar idea of "the stream of consciousness," and is initially deeply counterintuitive, but it grows on you, as you see how it handles facts about the brain that have been ignored up to now by philosophers — and scientists. By considering in some detail how consciousness could have evolved, we gain insights into otherwise baffling features of our minds. Part II also provides an analysis of the role of language in human consciousness, and the relation of the Multiple Drafts model to some more familiar conceptions of the mind, and to other theoretical work in the multidisciplinary field of cognitive science. All along the way we have to resist the alluring simplicities of the traditional view, until we can secure ourselves on the new foundation.

In Part III, armed with the new ways of guiding our imaginations, we can confront (at last) the traditional mysteries of consciousness: the strange properties of the "phenomenal field," the nature of introspection, the qualities (or *qualia*) of experiential states, the nature of the self or ego and its relation to thoughts and sensations, the consciousness of nonhuman creatures. The paradoxes that beset traditional philosophical debates about these can then be seen to arise from *failures of imagination*, not "insight," and we will be able to dissolve the mysteries.

This book presents a theory that is both empirical and philosophical, and since the demands on such a theory are so varied, there are two appendices that deal briefly with more technical challenges arising both from the scientific and philosophical perspectives. In the next chapter, we turn to the question of what an explanation of consciousness would be, and whether we should want to dissolve the mysteries of consciousness at all.

PART ONE

PROBLEMS AND METHODS

2

EXPLAINING CONSCIOUSNESS

1. PANDORA'S BOX: SHOULD CONSCIOUSNESS BE DEMYSTIFIED?

> And here are trees and I know their gnarled surface, water, and I feel its taste. These scents of grass and stars at night, certain evenings when the heart relaxes — how shall I negate this world whose power and strength I feel? Yet all the knowledge on earth will give me nothing to assure me that this world is mine. You describe it to me and you teach me to classify it. You enumerate its laws and in my thirst for knowledge I admit that they are true. You take apart its mechanism and my hope increases. . . . What need had I of so many efforts? The soft lines of these hills and the hand of evening on this troubled heart teach me much more.
>
> ALBERT CAMUS, *The Myth of Sisyphus,* 1942

> Sweet is the lore which Nature brings;
> Our meddling intellect
> Misshapes the beauteous forms of things: —
> We murder to dissect.
>
> WILLIAM WORDSWORTH, "The Tables Turned," 1798

Human consciousness is just about the last surviving mystery. A mystery is a phenomenon that people don't know how to think about — yet. There have been other great mysteries: the mystery of the origin of the universe, the mystery of life and reproduction, the mystery of the design to be found in nature, the mysteries of time, space, and gravity. These were not just areas of scientific ignorance, but of utter bafflement and wonder. We do not yet have the final answers to any of the questions of cosmology and particle physics, molecular genetics

and evolutionary theory, but we do know how to think about them. The mysteries haven't vanished, but they have been tamed. They no longer overwhelm our efforts to think about the phenomena, because now we know how to tell the misbegotten questions from the right questions, and even if we turn out to be dead wrong about some of the currently accepted answers, we know how to go about looking for better answers.

With consciousness, however, we are still in a terrible muddle. Consciousness stands alone today as a topic that often leaves even the most sophisticated thinkers tongue-tied and confused. And, as with all the earlier mysteries, there are many who insist — and hope — that there will never be a demystification of consciousness.

Mysteries are exciting, after all, part of what makes life fun. No one appreciates the spoilsport who reveals whodunit to the moviegoers waiting in line. Once the cat is out of the bag, you can never regain the state of delicious mystification that once enthralled you. So let the reader beware. If I succeed in my attempt to explain consciousness, those who read on will trade mystery for the rudiments of scientific knowledge of consciousness, not a fair trade for some tastes. Since some people view demystification as desecration, I expect them to view this book at the outset as an act of intellectual vandalism, an assault on the last sanctuary of humankind. I would like to change their minds.

Camus suggests he has no need of science, since he can learn more from the soft lines of the hills and the hand of evening, and I would not challenge his claim — given the questions Camus is asking himself. Science does not answer all good questions. Neither does philosophy. But for that very reason the phenomena of consciousness, which are puzzling in their own right quite independently of Camus's concerns, do not need to be protected from science — or from the sort of demystifying philosophical investigation we are embarking on. Sometimes people, fearing that science will "murder to dissect" as Wordsworth put it, are attracted to philosophical doctrines that offer one guarantee or another against such an invasion. The misgivings that motivate them are well founded, whatever the strengths and weaknesses of the doctrines; it indeed *could* happen that the demystification of consciousness would be a great loss. I will claim only that in fact this *will* not happen: the losses, if any, are overridden by the gains in understanding — both scientific and social, both theoretical and moral — that a good theory of consciousness can provide.

How, though, *might* the demystification of consciousness be something to regret? It might be like the loss of childhood innocence, which

is definitely a loss, even if it is well recompensed. Consider what happens to love, for instance, when we become more sophisticated. We can understand how a knight in the age of chivalry could want to sacrifice his life for the honor of a princess he had never so much as spoken to — this was an especially thrilling idea to me when I was about eleven or twelve — but it is not a state of mind into which an adult today can readily enter. People used to talk and think about love in ways that are now practically unavailable — except to children, and to those who can somehow suppress their adult knowledge. We all love to tell those we love that we love them, and to hear from them that we are loved — but as grownups we are not quite as sure we know what this means as we once were, when we were children and love was a simple thing.

Are we better or worse off for this shift in perspective? The shift is not uniform, of course. While naïve adults continue to raise gothic romances to the top of the best-seller list, we sophisticated readers find we have rendered ourselves quite immune to the intended effects of such books: they make us giggle, not cry. Or if they do make us cry — as sometimes they do, in spite of ourselves — we are embarrassed to discover that we are still susceptible to such cheap tricks; for we cannot readily share the mind-set of the heroine who wastes away worrying about whether she has found "true love" — as if this were some sort of distinct substance (emotional gold as opposed to emotional brass or copper). This growing up is not just in the individual. Our culture has become more sophisticated — or at least sophistication, whatever it is worth, is more widely spread through the culture. As a result, our concepts of love have changed, and with these changes come shifts in sensibility that now prevent us from having certain experiences that thrilled, devastated, or energized our ancestors.

Something similar is happening to consciousness. Today we talk about our conscious decisions and unconscious habits, about the conscious experiences we enjoy (in contrast to, say, automatic cash machines, which have no such experiences) — but we are no longer quite sure we know what we mean when we say these things. While there are still thinkers who gamely hold out for consciousness being some one genuine precious thing (like love, like gold), a thing that is just "obvious" and very, very special, the suspicion is growing that this is an illusion. Perhaps the various phenomena that conspire to create the sense of a single mysterious phenomenon have no more ultimate or essential unity than the various phenomena that contribute to the sense that love is a simple thing.

Compare love and consciousness with two rather different phenomena, diseases and earthquakes. Our concepts of diseases and earthquakes have also undergone substantial revision over the last few hundred years, but diseases and earthquakes are phenomena that are very largely (but not entirely) independent of our concepts of them. Changing our minds about diseases did not in itself make diseases disappear or become less frequent, although it did result in changes in medicine and public health that radically altered the occurrence patterns of diseases. Earthquakes may someday similarly come under some measure of human control, or at least prediction, but by and large the existence of earthquakes is unaffected by our attitudes toward them or concepts of them. With love it is otherwise. It is no longer possible for sophisticated people to "fall in love" in some of the ways that once were possible — simply because they cannot believe in those ways of falling in love. It is no longer possible for me, for instance, to have a pure teenaged crush — unless I "revert to adolescence" and in the process forget or abandon much of what I think I know. Fortunately, there are other kinds of love for me to believe in, but what if there weren't? Love is one of those phenomena that depend on their concepts, to put it oversimply for the time being. There are others; money is a clear instance. If everyone forgot what money was, there wouldn't be any money anymore; there would be stacks of engraved paper slips, embossed metal disks, computerized records of account balances, granite and marble bank buildings — but no money: no inflation or deflation or exchange rates or interest — or monetary value. The very property of those variously engraved slips of paper that explains — as nothing else could — their trajectories from hand to hand in the wake of various deeds and exchanges would evaporate.

On the view of consciousness I will develop in this book, it turns out that consciousness, like love and money, is a phenomenon that does indeed depend to a surprising extent on its associated concepts. Although, like love, it has an elaborate biological base, like money, some of its most significant features are borne along on the culture, not simply inherent, somehow, in the physical structure of its instances. So if I am right, and if I succeed in overthrowing some of those concepts, I will threaten with extinction whatever phenomena of consciousness depend on them. Are we about to enter the postconscious period of human conceptualization? Is this not something to fear? Is it even conceivable?

If the concept of consciousness were to "fall to science," what would happen to our sense of moral agency and free will? If conscious

experience were "reduced" somehow to mere matter in motion, what would happen to our appreciation of love and pain and dreams and joy? If conscious human beings were "just" animated material objects, how could anything we do to them be right or wrong? These are among the fears that fuel the resistance and distract the concentration of those who are confronted with attempts to explain consciousness.

I am confident that these fears are misguided, but they are not obviously misguided. They raise the stakes in the confrontation of theory and argument that is about to begin. There are powerful arguments, quite independent of the fears, arrayed against the sort of scientific, materialistic theory I will propose, and I acknowledge that it falls to me to demonstrate not only that these arguments are mistaken, but also that the widespread acceptance of my vision of consciousness would not have these dire consequences in any case. (And if I had discovered that it would likely have these effects — what would I have done then? I wouldn't have written *this* book, but beyond that, I just don't know.)

Looking on the bright side, let us remind ourselves of what has happened in the wake of earlier demystifications. We find no diminution of wonder; on the contrary, we find deeper beauties and more dazzling visions of the complexity of the universe than the protectors of mystery ever conceived. The "magic" of earlier visions was, for the most part, a cover-up for frank failures of imagination, a boring dodge enshrined in the concept of a *deus ex machina*. Fiery gods driving golden chariots across the skies are simpleminded comic-book fare compared to the ravishing strangeness of contemporary cosmology, and the recursive intricacies of the reproductive machinery of DNA make *élan vital* about as interesting as Superman's dread kryptonite. When we understand consciousness — when there is no more mystery — consciousness will be different, but there will still be beauty, and more room than ever for awe.

2. THE MYSTERY OF CONSCIOUSNESS

What, then, is the mystery? What could be more obvious or certain to each of us than that he or she is a conscious subject of experience, an enjoyer of perceptions and sensations, a sufferer of pain, an entertainer of ideas, and a conscious deliberator? That seems undeniable, but what in the world can consciousness itself be? How can living physical bodies in the physical world produce such phenomena? That is the mystery.

The mystery of consciousness has many ways of introducing itself, and it struck me anew with particular force one recent morning as I sat in a rocking chair reading a book. I had apparently just looked up from my book, and at first had been gazing blindly out the window, lost in thought, when the beauty of my surroundings distracted me from my theoretical musings. Green-golden sunlight was streaming in the window that early spring day, and the thousands of branches and twigs of the maple tree in the yard were still clearly visible through a mist of green buds, forming an elegant pattern of wonderful intricacy. The windowpane is made of old glass, and has a scarcely detectable wrinkle line in it, and as I rocked back and forth, this imperfection in the glass caused a wave of synchronized wiggles to march back and forth across the delta of branches, a regular motion superimposed with remarkable vividness on the more chaotic shimmer of the twigs and branches in the breeze.

Then I noticed that this visual metronome in the tree branches was locked in rhythm with the Vivaldi concerto grosso I was listening to as "background music" for my reading. At first I thought it was obvious that I must have unconsciously synchronized my rocking with the music — just as one may unconsciously tap one's foot in time — but rocking chairs actually have a rather limited range of easily maintained rocking frequencies, so probably the synchrony was mainly a coincidence, just slightly pruned by some unconscious preference of mine for neatness, for staying in step.

In my mind I skipped fleetingly over some dimly imagined brain processes that might explain how we unconsciously adjust our behavior, including the behavior of our eyes and our attention-directing faculties, in order to "synchronize" the "sound track" with the "picture," but these musings were interrupted in turn by an abrupt realization. *What I was doing* — the interplay of experiencing and thinking I have just described from my privileged, first-person point of view — was much harder to "make a model of" than the unconscious, backstage processes that were no doubt going on in me and somehow the causal conditions for what I was doing. Backstage machinery was relatively easy to make sense of; it was the front-and-center, in-the-limelight goings-on that were downright baffling. My conscious thinking, and especially the enjoyment I felt in the combination of sunny light, sunny Vivaldi violins, rippling branches — plus the pleasure I took in just thinking about it all — how could *all that* be just something physical happening in my brain? How could any combination of electrochemical happenings in my brain somehow add up to the delightful way those

hundreds of twigs genuflected in time with the music? How could some information-processing event in my brain be the delicate warmth of the sunlight I felt falling on me? For that matter, how could an event in my brain be my sketchily visualized mental image of . . . some other information-processing event in my brain? It does seem impossible.

It does seem as if the happenings that are my conscious thoughts and experiences cannot be brain happenings, but must be *something else*, something caused or produced by brain happenings, no doubt, but something in addition, made of different stuff, located in a different space. Well, why not?

3. THE ATTRACTIONS OF MIND STUFF

Let's see what happens when we take this undeniably tempting route. First, I want you to perform a simple experiment. It involves closing your eyes, imagining something, and then, once you have formed your mental image and checked it out carefully, answering some questions below. Do not read the questions until after you have followed this instruction: when you close your eyes, imagine, in as much detail as possible, a purple cow.

Done? Now:

(1) Was your cow facing left or right or head on?
(2) Was she chewing her cud?
(3) Was her udder visible to you?
(4) Was she a relatively pale purple, or deep purple?

If you followed instructions, you could probably answer all four questions without having to make something up in retrospect. If you found all four questions embarrassingly demanding, you probably didn't bother imagining a purple cow at all, but just thought, lazily: "I'm imagining a purple cow" or "Call *this* imagining a purple cow," or did something nondescript of that sort.

Now let us do a second exercise: close your eyes and imagine, in as much detail as possible, a *yellow* cow.

This time you can probably answer the first three questions above without any qualms, and will have something confident to say about what sort of yellow — pastel or buttery or tan — covered the flanks of your imagined cow. But this time I want to consider a different question:

(5) What is the difference between imagining a purple cow and imagining a yellow cow?

The answer is obvious: The first imagined cow is purple and the second is yellow. There might be other differences, but that is the essential one. The trouble is that since these cows are just imagined cows, rather than real cows, or painted pictures of cows on canvas, or cow shapes on a color television screen, it is hard to see what could be purple in the first instance and yellow in the second. Nothing roughly cow-shaped in your brain (or in your eyeball) turns purple in one case and yellow in the other, and even if it did, this would not be much help, since it's pitch black inside your skull and, besides, you haven't any eyes in there to see colors with.

There are events in your brain that are tightly associated with your particular imaginings, so it is not out of the question that in the near future a neuroscientist, examining the processes that occurred in your brain in response to my instructions, would be able to decipher them to the extent of being able to confirm or disconfirm your answers to questions 1 through 4:

"Was the cow facing left? We think so. The cow-head neuronal excitation pattern was consistent with upper-left visual quadrant presentation, and we observed one-herz oscillatory motion-detection signals that suggest cud-chewing, but we could detect no activity in the udder-complex representation groups, and, after calibration of evoked potentials with the subject's color-detection profiles, we hypothesize that the subject is lying about the color: the imagined cow was almost certainly brown."

Suppose all this were true; suppose scientific mind-reading had come of age. Still, it seems, the mystery would remain: what is brown when you imagine a brown cow? Not the event in the brain that the scientists have calibrated with your experiencing-of-brown. The type and location of the neurons involved, their connections with other parts of the brain, the frequency or amplitude of activity, the neurotransmitter chemicals released — none of those properties is the very property of the cow "in your imagination." And since you did imagine a cow (you are not lying — the scientists even confirm that), an imagined cow came into existence at that time; something, somewhere must have had those properties at that time. The imagined cow must be rendered not in the medium of brain stuff, but in the medium of . . . mind stuff. What else could it be?

Mind stuff, then, must be "what dreams are made of," and it apparently has some remarkable properties. One of these we have already noticed in passing, but it is extremely resistant to definition. As a first pass, let us say that mind stuff always *has a witness*. The trouble

with brain events, we noticed, is that no matter how closely they "match" the events in our streams of consciousness, they have one apparently fatal drawback: *There's nobody in there watching them.* Events that happen in your brain, just like events that happen in your stomach or your liver, are not normally witnessed by anyone, nor does it make any difference to how they happen whether they occur witnessed or unwitnessed. Events in consciousness, on the other hand, are "by definition" witnessed; they are experienced by an experiencer, and their being thus experienced is what makes them what they are: conscious events. An experienced event cannot just happen on its own hook, it seems; it must be somebody's experience. For a thought to happen, someone (some mind) must think it, and for a pain to happen, someone must feel it, and for a purple cow to burst into existence "in imagination," someone must imagine it.

And the trouble with brains, it seems, is that when you look in them, you discover that there's nobody home. No part of the brain is the thinker that does the thinking or the feeler that does the feeling, and the whole brain appears to be no better a candidate for that very special role. This is a slippery topic. Do brains think? Do eyes see? Or do people see with their eyes and think with their brains? Is there a difference? Is this just a trivial point of "grammar" or does it reveal a major source of confusion? The idea that a *self* (or a person, or, for that matter, a soul) is distinct from a brain or a body is deeply rooted in our ways of speaking, and hence in our ways of thinking.

I have a brain.

This seems to be a perfectly uncontroversial thing to say. And it does not seem to mean just

This body has a brain (and a heart, and two lungs, etc.).

or

This brain has itself.

It is quite natural to think of "the self and its brain" (Popper and Eccles, 1977) as two distinct things, with different properties, no matter how closely they depend on each other. If the self is distinct from the brain, it seems that it must be made of mind stuff. In Latin, a thinking thing is a *res cogitans*, a term made famous by Descartes, who offered what he thought was an unshakable proof that he, manifestly a thinking thing, could not be his brain. Here is one of his versions of it, and it is certainly compelling:

> I next considered attentively what I was; and I saw that while I could pretend that I had no body, that there was no world, and no place for me to be in, I could not pretend that I was not; on the contrary, from the mere fact that I thought of doubting the truth of other things it evidently and certainly followed that I existed. On the other hand, if I had merely ceased to think, even if everything else that I had ever imagined had been true, I had no reason to believe that I should have existed. From this I recognized that I was a substance whose whole essence or nature is to think and whose being requires no place and depends on no material thing. [*Discourse on Method*, 1637]

So we have discovered two sorts of things one might want to make out of mind stuff: the purple cow that isn't in the brain, and the thing that does the thinking. But there are still other special powers we might want to attribute to mind stuff.

Suppose a winery decided to replace their human wine tasters with a machine. A computer-based "expert system" for quality control and classification of wine is almost within the bounds of existing technology. We now know enough about the relevant chemistry to make the transducers that would replace the taste buds and the olfactory receptors of the epithelium (providing the "raw material" — the input stimuli — for taste and smell). How these inputs combine and interact to produce our experiences is not precisely known, but progress is being made. Work on vision has proceeded much farther. Research on color vision suggests that mimicking human idiosyncrasy, delicacy, and reliability in the color-judging component of the machine would be a great technical challenge, but it is not out of the question. So we can readily imagine using the advanced outputs of these sensory transducers and their comparison machinery to feed elaborate classification, description, and evaluation routines. Pour the sample wine in the funnel and, in a few minutes or hours, the system would type out a chemical assay, along with commentary: "a flamboyant and velvety Pinot, though lacking in stamina" — or words to such effect. Such a machine might even perform better than human wine tasters on all reasonable tests of accuracy and consistency the winemakers could devise, but surely no matter how "sensitive" and "discriminating" such a system might become, it seems that it would never have, and enjoy, what *we* do when we taste a wine.

Is this in fact so obvious? According to the various ideologies

grouped under the label of *functionalism*, if you reproduced the entire "functional structure" of the human wine taster's cognitive system (including memory, goals, innate aversions, etc.), you would thereby reproduce *all* the mental properties as well, including the enjoyment, the delight, the savoring that makes wine-drinking something many of us *appreciate*. In principle it makes no difference, the functionalist says, whether a system is made of organic molecules or silicon, so long as it does *the same job*. Artificial hearts don't have to be made of organic tissue, and neither do artificial brains — at least in principle. If all the control functions of a human wine taster's brain can be reproduced in silicon chips, the enjoyment will *ipso facto* be reproduced as well.

Some brand of functionalism may triumph in the end (in fact this book will defend a version of functionalism), but it surely seems outrageous at first blush. It seems that no mere machine, no matter how accurately it mimicked the brain processes of the human wine taster, would be capable of appreciating a wine, or a Beethoven sonata, or a basketball game. For appreciation, you need consciousness — something no mere machine has. But of course the brain *is* a machine of sorts, an organ like the heart or lungs or kidneys with an ultimately mechanical explanation of all its powers. This can make it seem compelling that the brain isn't what does the appreciating; *that* is the responsibility (or privilege) of the mind. Reproduction of the brain's machinery in a silicon-based machine wouldn't, then, yield real appreciation, but at best the illusion or simulacrum of appreciation.

So the conscious mind is not just the place where the witnessed colors and smells are, and not just the thinking thing. It is where the appreciating happens. It is the ultimate arbiter of why anything matters. Perhaps this even follows somehow from the fact that the conscious mind is also supposed to be the source of our intentional actions. It stands to reason — doesn't it? — that if *doing things that matter* depends on consciousness, *mattering* (enjoying, appreciating, suffering, caring) should depend on consciousness as well. If a sleepwalker "unconsciously" does harm, he is not responsible because in an important sense he didn't do it; his bodily motions are intricately involved in the causal chains that led to the harm, but they did not constitute any *actions* of his, any more than if he had simply done the harm by falling out of bed. Mere bodily complicity does not make for an intentional action, nor does bodily complicity under the control *of structures in*

the brain, for a sleepwalker's body is manifestly under the control of structures in the sleepwalker's brain. What more must be added is consciousness, the special ingredient that turns mere *happenings* into *doings*.[1]

It is not Vesuvius's fault if its eruption kills your beloved, and resenting (Strawson, 1962) or despising it are not available options — unless you somehow convince yourself that Vesuvius, contrary to contemporary opinion, is a conscious agent. It is indeed strangely comforting in our grief to put ourselves into such states of mind, to rail at the "fury" of the hurricane, to curse the cancer that so unjustly strikes down a child, or to curse "the gods." Originally, to say that something was "animate" as opposed to "inanimate" was to say that it had a soul (*anima* in Latin). It may be more than just comforting to think of the things that affect us powerfully as animate; it may be a deep biological design trick, a shortcut for helping our time-pressured brains organize and think about the things that need thinking about if we are to survive.

We might have an innate tendency to treat every changing thing at first as if it had a soul (Stafford, 1983; Humphrey, 1983b, 1986), but however natural this attitude is, we now know that attributing a (conscious) soul to Vesuvius is going too far. Just where to draw the line is a vexing question to which we will return, but for ourselves, it seems, consciousness is precisely what distinguishes us from mere "automata." Mere bodily "reflexes" are "automatic" and mechanical; they may involve circuits in the brain, but do not require any intervention by the conscious mind. It is very natural to think of our own bodies as mere hand puppets of sorts that "we" control "from inside." I make the hand puppet wave to the audience by wiggling my finger; I wiggle my finger by . . . what, wiggling my soul? There are notorious problems with this idea, but that does not prevent it from seeming somehow right: unless there is a conscious mind behind the deed, there is no *real* agent in charge. When we think of our minds this way, we seem to discover the "inner me," the "real me." This real me is not my brain; it is what *owns* my brain ("the self and *its* brain"). On Harry Truman's desk in the Oval Office of the White House was a famous sign: "The buck stops here." No part of the brain, it seems, could be *where the buck stops*, the ultimate source of moral responsibility at the beginning of a chain of command.

To summarize, we have found four reasons for believing in mind

1. See my *Elbow Room* (1984), chapter 4, for further discussion of this theme.

stuff. The conscious mind, it seems, cannot just be the brain, or any proper part of it, because nothing in the brain could

(1) be the medium in which the purple cow is rendered;
(2) be the thinking thing, the *I* in "I think, therefore I am";
(3) appreciate wine, hate racism, love someone, be a source of *mattering;*
(4) act with moral responsibility.

An acceptable theory of human consciousness must account for these four compelling grounds for thinking that there must be mind stuff.

4. WHY DUALISM IS FORLORN

The idea of mind as distinct in this way from the brain, composed not of ordinary matter but of some other, special kind of stuff, is *dualism,* and it is deservedly in disrepute today, in spite of the persuasive themes just canvassed. Ever since Gilbert Ryle's classic attack (1949) on what he called Descartes's "dogma of the ghost in the machine," dualists have been on the defensive.[2] The prevailing wisdom, variously expressed and argued for, is *materialism:* there is only one sort of stuff, namely *matter* — the physical stuff of physics, chemistry, and physiology — and the mind is somehow nothing but a physical phenomenon. In short, the mind is the brain. According to the materialists, we can (in principle!) account for every mental phenomenon using the same physical principles, laws, and raw materials that suffice to explain radioactivity, continental drift, photosynthesis, reproduction, nutrition, and growth. It is one of the main burdens of this book to explain consciousness without ever giving in to the siren song of dualism. What, then, is so wrong with dualism? Why is it in such disfavor?

The standard objection to dualism was all too familiar to Descartes himself in the seventeenth century, and it is fair to say that neither he nor any subsequent dualist has ever overcome it convincingly. If mind

2. A few brave souls (and they surely cannot object to being so categorized!) have bucked the tide: Arthur Koestler's defiantly titled *The Ghost in the Machine* (1967) and Popper and Eccles's *The Self and Its Brain* (1977) are by unquestionably eminent authors, and two other iconoclastic and quirkily insightful defenses of dualism are Zeno Vendler's *Res Cogitans* (1972) and *The Matter of Minds* (1984).

and body are distinct things or substances, they nevertheless must interact; the bodily sense organs, via the brain, must *inform* the mind, must send to it or present it with perceptions or ideas or data of some sort, and then the mind, having thought things over, must *direct* the body in appropriate action (including speech). Hence the view is often called Cartesian interactionism or interactionist dualism. In Descartes's formulation, the locus of interaction in the brain was the pineal gland, or *epiphysis*. It appears in Descartes's own schematic diagram as the much-enlarged pointed oval in the middle of the head.

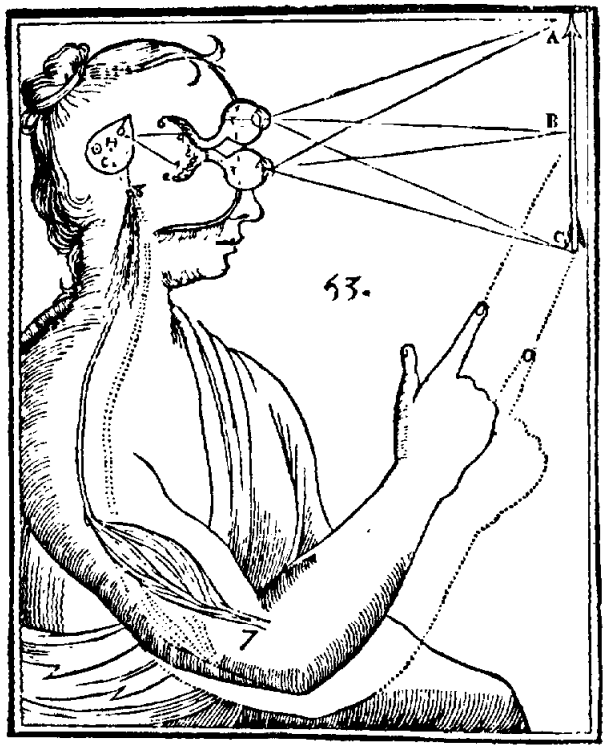


Figure 2.1

We can make the problem with interactionism clear by superimposing a sketch of the rest of Descartes's theory on his diagram (Figure 2.2).

The conscious perception of the arrow occurs only after the brain has somehow transmitted its message to the mind, and the person's finger can point to the arrow only after the mind commands the body. How, precisely, does the information get transmitted from pineal gland to mind? Since we don't have the faintest idea (yet) what properties mind stuff has, we can't even guess (yet) how it might be affected by physical processes emanating somehow from the brain, so let's ignore those upbound signals for the time being, and concentrate on the return signals, the directives from mind to brain. These, ex *hypothesi*, are not physical; they are not light waves or sound waves or cosmic rays or

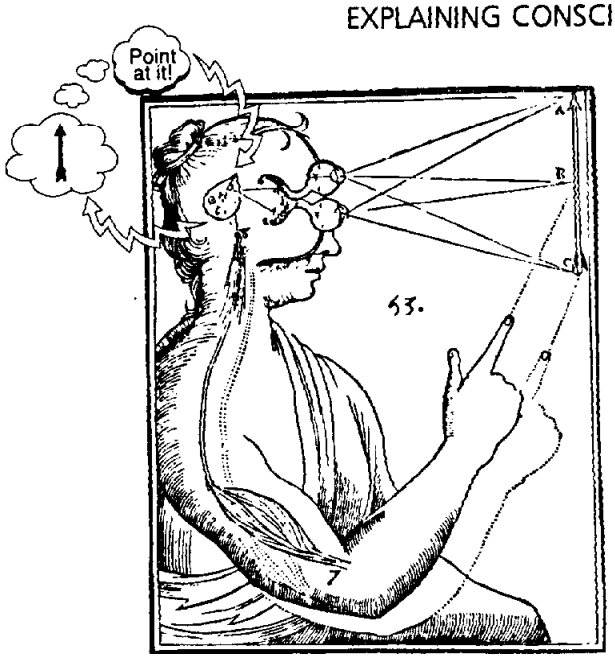


Figure 2.2

streams of subatomic particles. No physical energy or mass is associated with them. How, then, do they get to make a difference to what happens in the brain cells they must affect, if the mind is to have any influence over the body? A fundamental principle of physics is that any change in the trajectory of any physical entity is an acceleration requiring the expenditure of energy, and where is this energy to come from? It is this principle of the conservation of energy that accounts for the physical impossibility of "perpetual motion machines," and the same principle is apparently violated by dualism. This confrontation between quite standard physics and dualism has been endlessly discussed since Descartes's own day, and is widely regarded as the inescapable and fatal flaw of dualism.

Just as one would expect, ingenious technical exemptions based on sophisticated readings of the relevant physics have been explored and expounded, but without attracting many conversions. Dualism's embarrassment here is really simpler than the citation of presumed laws of physics suggests. It is the same incoherence that children notice — but tolerate happily in fantasy — in such fare as Casper the Friendly Ghost (Figure 2.3, page 36). How can Casper *both* glide through walls and grab a falling towel? How can mind stuff *both* elude all physical measurement and control the body? A ghost in the machine is of no help in our theories unless it is a ghost that can move things around — like a noisy poltergeist who can tip over a lamp or slam a door — but anything that can move a physical thing is itself a physical thing (although perhaps a strange and heretofore unstudied kind of physical thing).

What about the option, then, of concluding that mind stuff is

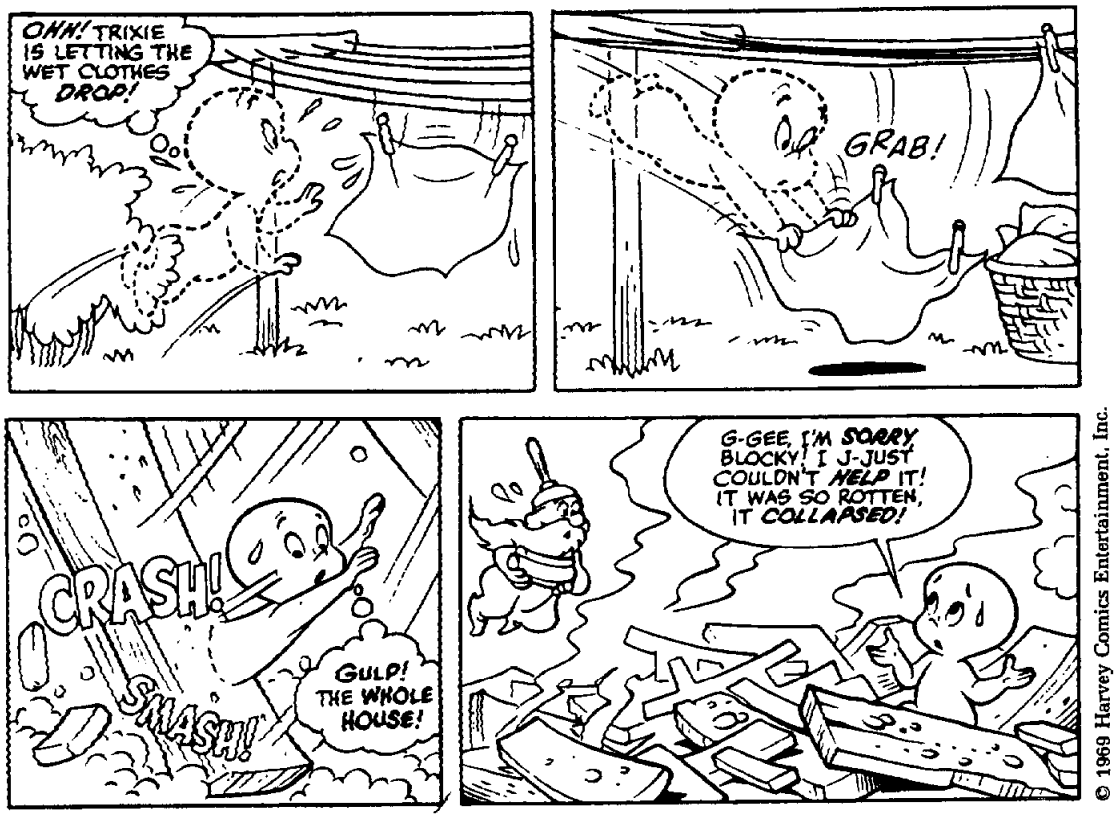


Figure 2.3

actually a special kind of matter? In Victorian séances, the mediums often produced out of thin air something they called "ectoplasm," a strange gooey substance that was supposedly the basic material of the spirit world, but which could be trapped in a glass jar, and which oozed and moistened and reflected light just like everyday matter. Those fraudulent trappings should not dissuade us from asking, more soberly, whether mind stuff might indeed be something above and beyond the atoms and molecules that compose the brain, but still a scientifically investigatable kind of matter. The ontology of a theory is the catalogue of things and types of things the theory deems to exist. The ontology of the physical sciences used to include "caloric" (the stuff heat was made of, in effect) and "the ether" (the stuff that pervaded space and was the medium of light vibrations in the same way air or water can be the medium of sound vibrations). These things are no longer taken seriously, while neutrinos and antimatter and black holes are now included in the standard scientific ontology. Perhaps some basic enlargement of the ontology of the physical sciences is called for in order to account for the phenomena of consciousness.

Just such a revolution of physics has recently been proposed by the physicist and mathematician Roger Penrose, in *The Emperor's New Mind* (1989). While I myself do not think he has succeeded in making

his case for revolution,[3] it is important to notice that he has been careful *not* to fall into the trap of dualism. What is the difference? Penrose makes it clear that he intends his proposed revolution to make the conscious mind more accessible to scientific investigation, not less. It is surely no accident that the few dualists to avow their views openly have all candidly and comfortably announced that they have no theory whatever of how the mind works — something, they insist, that is quite beyond human ken.[4] There is the lurking suspicion that the most attractive feature of mind stuff is its promise of being so mysterious that it keeps science at bay forever.

This fundamentally antiscientific stance of dualism is, to my mind, its most disqualifying feature, and is the reason why in this book I adopt the apparently dogmatic rule that dualism is to be avoided *at all costs*. It is not that I think I can give a knock-down proof that dualism, in all its forms, is false or incoherent, but that, given the way dualism wallows in mystery, *accepting dualism is giving up* (as in Figure 2.4, page 38).

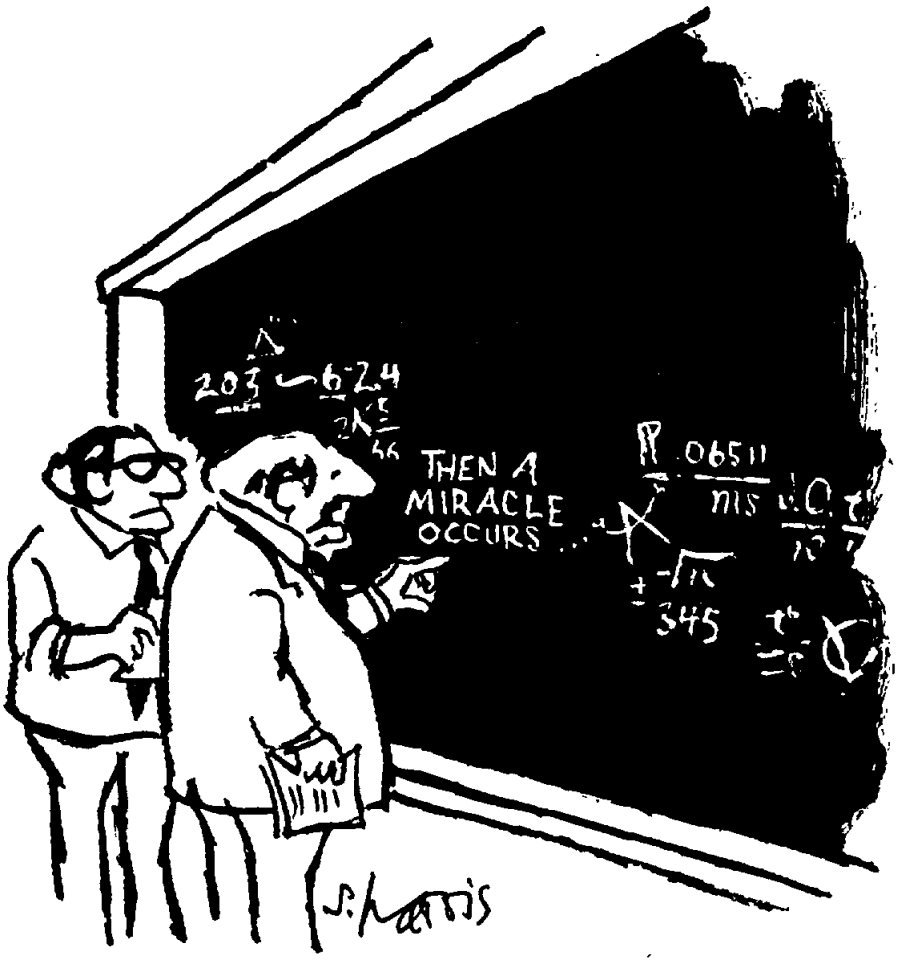
There is widespread agreement about this, but it is as shallow as it is wide, papering over some troublesome cracks in the materialist wall. Scientists and philosophers may have achieved a consensus of sorts in favor of materialism, but as we shall see, getting rid of the old dualistic visions is harder than contemporary materialists have thought. Finding suitable replacements for the traditional dualistic images will require some rather startling adjustments to our habitual ways of thinking, adjustments that will be just as counterintuitive at first to scientists as to laypeople.

I don't view it as ominous that my theory seems at first to be strongly at odds with common wisdom. On the contrary, we shouldn't expect a good theory of consciousness to make for comfortable reading — the sort that immediately "rings bells," that makes us exclaim to ourselves, with something like secret pride: "Of course! I knew that all along! It's obvious, once it's been pointed out!" If there were any such theory to be had, we would surely have hit upon it by now. The mysteries of the mind have been around for so long, and we have made

---

3. See "Murmurs in the Cathedral" (Dennett, 1989c), my review of his book.

4. Eccles has proposed that the nonphysical mind is composed of millions of "psychons," which interact with millions of "dendrons" (tracts of pyramidal cells) in the cortex; each psychon corresponds roughly to what Descartes or Hume would call an idea — such as the idea of red, or the idea of round, or hot — but aside from this minimal decomposition, Eccles has nothing to say about the parts, activities, principles of action, or other properties of the nonphysical mind.



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Figure 2.4

so little progress on them, that the likelihood is high that some things we all tend to agree to be obvious are just not so. I will soon be introducing my candidates.

Some brain researchers today — perhaps even a stolid majority of them — continue to pretend that, for them, the brain is just another organ, like the kidney or pancreas, which should be described and explained only in the most secure terms of the physical and biological sciences. They would never dream of mentioning the mind or anything "mental" in the course of their professional duties. For other, more theoretically daring researchers, there is a new object of study, the mind/brain (Churchland, 1986). This newly popular coinage nicely expresses the prevailing materialism of these researchers, who happily admit to

the world and themselves that what makes the brain particularly fascinating and baffling is that somehow or other it *is* the mind. But even among these researchers there is a reluctance to confront the Big Issues, a desire to postpone until some later date the embarrassing questions about the nature of consciousness.

But while this attitude is entirely reasonable, a modest recognition of the value of the divide-and-conquer strategy, it has the effect of distorting some of the new concepts that have arisen in what is now called cognitive science. Almost all researchers in cognitive science, whether they consider themselves neuroscientists or psychologists or artificial intelligence researchers, tend to postpone questions about consciousness by restricting their attention to the "peripheral" and "subordinate" systems of the mind/brain, which are deemed to feed and service some dimly imagined "center" where "conscious thought" and "experience" take place. This tends to have the effect of leaving too much of the mind's work to be done "in the center," and this leads theorists to underestimate the "amount of understanding" that must be accomplished by the relatively peripheral systems of the brain (Dennett, 1984b).

For instance, theorists tend to think of perceptual systems as providing "input" to some central thinking arena, which in turn provides "control" or "direction" to some relatively peripheral systems governing bodily motion. This central arena is also thought to avail itself of material held in various relatively subservient systems of memory. But the very idea that there are important theoretical divisions between such presumed subsystems as "long-term memory" and "reasoning" (or "planning") is more an artifact of the divide-and-conquer strategy than anything found in nature. As we shall soon see, the exclusive attention to specific subsystems of the mind/brain often causes a sort of theoretical myopia that prevents theorists from seeing that their models still presuppose that somewhere, conveniently hidden in the obscure "center" of the mind/brain, there is a Cartesian Theater, a place where "it all comes together" and consciousness happens. This may seem like a good idea, an inevitable idea, but until we see, in some detail, why it is not, the Cartesian Theater will continue to attract crowds of theorists transfixed by an illusion.

5. THE CHALLENGE

In the preceding section, I noted that if dualism is the best we can do, then we can't understand human consciousness. Some people are

convinced that we can't in any case. Such defeatism, today, in the midst of a cornucopia of scientific advances ready to be exploited, strikes me as ludicrous, even pathetic, but I suppose it could be the sad truth. Perhaps consciousness really can't be explained, but how will we know till someone tries? I think that many — indeed, most — of the pieces of the puzzle are already well understood, and only need to be jiggled into place with a little help from me. Those who would defend the Mind against Science should wish me luck with this attempt, since if they are right, my project is bound to fail, but if I do the job about as well as it could be done, my failure ought to shed light on just why science will always fall short. They will at last have their argument against science, and I will have done all the dirty work for them.

The ground rules for my project are straightforward:

(1) *No Wonder Tissue allowed.* I will try to explain every puzzling feature of human consciousness within the framework of contemporary physical science; at no point will I make an appeal to inexplicable or unknown forces, substances, or organic powers. In other words, I intend to see what can be done within the conservative limits of standard science, saving a call for a revolution in materialism as a last resort.

(2) *No feigning anesthesia.* It has been said of behaviorists that they feign anesthesia — they pretend they don't have the experiences we know darn well they share with us. If I wish to deny the existence of some controversial feature of consciousness, the burden falls on me to *show* that it is somehow illusory.

(3) *No nitpicking about empirical details.* I will try to get all the scientific facts right, insofar as they are known today, but there is abundant controversy about just which exciting advances will stand the test of time. If I were to restrict myself to "facts that have made it into the textbooks," I would be unable to avail myself of some of the most eye-opening recent discoveries (if that is what they are). And I would *still* end up unwittingly purveying some falsehoods, if recent history is any guide. Some of the "discoveries" about vision for which David Hubel and Torstein Wiesel were deservedly awarded the Nobel Prize in 1981 are now becoming unraveled, and Edwin Land's famous "retinex" theory of color vision, which has been regarded by most philosophers of mind and other

nonspecialists as established fact for more than twenty years, is not nearly as highly regarded among visual scientists.[5]

So, since as a philosopher I am concerned to establish the *possibilities* (and rebut claims of impossibility), I will settle for theory sketches instead of full-blown, empirically confirmed theories. A theory sketch or a model of how the brain might do something can turn a perplexity into a research program: if this model won't quite do, would some other more realistic variation do the trick? (The explanation sketch of hallucination production in chapter 1 is an example of this.) Such a sketch is directly and explicitly vulnerable to empirical disproof, but if you want to claim that my sketch is not a *possible* explanation of a phenomenon, you must show what it *has to* leave out or *cannot* do; if you merely claim that my model may well be incorrect in many of its details, I will concede the point. What is wrong with Cartesian dualism, for instance, is not that Descartes chose the pineal gland — as opposed to the thalamus, say, or the amygdala — as the locus of interaction with the mind, but the very *idea* of such a locus of mind-brain interaction. What counts as nitpicking changes, of course, as science advances, and different theorists have different standards. I will try to err on the side of overspecificity, not only to heighten the contrast with traditional philosophy of mind, but to give empirical critics a clearer target at which to shoot.

In this chapter, we have encountered the basic features of the mystery of consciousness. The very mysteriousness of consciousness is one of its central features — possibly even a vital feature without which it cannot survive. Since this possibility is widely if dimly appreciated, prudence tends to favor doctrines that do not even purport to explain consciousness, for consciousness matters deeply to us. Dualism, the idea that a brain cannot be a thinking thing so a thinking thing cannot be a brain, is tempting for a variety of reasons, but we must resist temptation; "adopting" dualism is really just accepting defeat without admitting it. Adopting materialism does not by itself dissolve the puzzles about consciousness, nor do they fall to any straightforward inferences from brain science. Somehow the brain must be the mind, but unless we can come to see in some detail how this is possible, our

5. A fascinating review of the status of Land's theory is provided by the philosopher C. L. Hardin in an appendix to his book, *Color for Philosophers: Unweaving the Rainbow* (1988).

materialism will not explain consciousness, but only promise to explain it, some sweet day. That promise cannot be kept, I have suggested, until we learn how to abandon more of Descartes's legacy. At the same time, whatever else our materialist theories may explain, they won't explain consciousness if we neglect the facts about experience that we know so intimately "from the inside." In the next chapter, we will develop an initial inventory of those facts.