sufficiently developed that it becomes aware of sensory stimuli such as sounds. Further, it is uncertain how we should think of conscious states such as recognizing that something is unfamiliar or odd, or that something is intellectually satisfying, morally unsettling, musically harmonious, or esthetically jarring.

Fortunately, we need not worry too much at this stage about these cases. By identifying prototypical examples of conscious states, we gain lots of scope for designing revealing, interpretable experiments. With some progress in hand, less central examples may come to assume greater importance, perhaps even gain recognition as *the* prototypical cases.

Cognizant of the possibility that these ostensibly obvious categories may be reconfigured later under the pressure of new discoveries, perhaps we can agree that this rough-and-ready delineation of prototypes provides us with a reasonable way to get the project off the ground. Because the neuroscientific approach to consciousness is young, the reasonable hope is for discoveries that will open more doors and suggest fruitful experimental research. In the long haul, of course, we want to understand consciousness at least as well as we understand reproduction or metabolism, but in the short haul, it is wise to have realistic goals. It is probably not realistic to expect, for example, that a single experimental paradigm will solve the mystery.

1.3 Experimental Strategies

Although there are many proposals for making progress experimentally, for convenience the strategies targeting the brain can roughly be grouped as one of two kinds: a *direct approach* or an *indirect approach*. These strategies differ mainly in emphasis. In any case, as will be seen, they are *complementary*, not mutually incompatible. To see the strengths and weaknesses of each, I shall outline the somewhat differing motivations, scientific styles, and experimental approaches.

The direct approach

It is possible, for all we can tell now, that consciousness, or at least the sensory component of consciousness, may be subserved by a physical substrate with a distinctive signature. In the hope that there is some distinct and discernible physical marker of the substrate, the direct strategy aims first to identify the substrate as a *correlate* of phenomenological awareness, then eventually to get a reductive explanation of conscious states in neurobiological terms. The physical substrate need not be confined to one location. It could, for example, consist in a pattern of activity in one or two structurally unique cell *types* found in a particular layer of cortex across a range of brain areas. Or it could consist in the synchronized firing of a special cell population in the thalamus and certain cortical areas. On these alternatives, the mechanism would be *distributed*, and hence would be more like the endocrine system, for example, than the kidney. For convenience, I shall refer to a postulated physical substrate as a *mechanism* for consciousness.

Notice also that the distinctive mechanism could reside at any of a variety of physical levels: molecular, single cell, circuit, pathway, or some higher organizational level not yet explicitly catalogued. Or perhaps consciousness is the product of interactions between these myriad physical levels. The possibility of a distributed mechanism, together with the opened-ended possibility concerning the *level* of organization at which the mechanism inheres, means that hypotheses are so far quite unconstrained. The lack of constraints is not a symptom of anything otherworldly about this problem. It is merely a symptom that science has a lot of work to do.

Discovering some one or more of the neural correlates of consciousness would not *on its own* yield an explanation of consciousness. Nevertheless, in biology the discovery of which mechanism supports a specific function often means that the next step—determining precisely *how* the function is performed—suddenly becomes a whole lot easier. Not *easy*, but easier. Were we lucky enough to identify the hypothetical mechanism, the result would be comparable in its scientific ramifications to identifying the structure of DNA. That discovery was essentially a discovery about structural embodiment of information. Once the structure of the double helix was revealed, it became possible to see that the order of the base pairs was a code for making proteins, and hence to understand the structural basis for heritability of traits. In the event that there is a mechanism with a distinct signature identifiable with conscious states, the scientific payoff *could* be enormous. The direct strategy, therefore, is worth a good shot.

The downside, of course, is that the mechanism might be experimentally very difficult to identify until neuroscience is *much* further along, since the signature may not be obvious to the naive observer. Our current misconceptions about the phenomena to be explained, or about the brain, may lead us to misinterpret the data even if the mechanism with its distinct signature exists to be identified. Or there may be other unforeseeable pitfalls to bedevil the approach. In short, all the usual problems besetting any ambitious scientific project beset us here.

In recent years, the direct approach has become more clearly articulated and more experimentally attractive, in part occasioned by new techniques that made it possible to investigate closely related functions such as attention and working memory.

Francis Crick, probably more than anyone else, has a sure-footed scientific sense of what the direct approach would need to succeed. He has drawn attention to the value of using low-level and systems-level data to narrow the search space of plausible hypotheses, and of constantly prowling that search space to provoke one's scientific imagination to come up with testable hypotheses. Crick has consistently recognized and defended the value of getting *some* sort of structural bead on the neuroanatomy subserving conscious states, not because he thought such data would solve the problems in one grand sweep, but because he realized it would give us a thread, which, when pulled, might begin to unravel the problem. He argued that experiments probing such a mechanism could make a plausible assumption, which I henceforth refer to as Crick's assumption:

Crick's assumption There must be brain differences in the following two conditions: (1) a stimulus is presented and the subject *is* aware of it, and (2) a stimulus is presented and the subject is *not* aware of it.[5]

With the right experiments, it should be possible to find what is different about the brain in these two conditions.

Within this lean framework, the next step is to find an experimental paradigm where psychology and neuroscience can hold hands across the divide; in other words, to find a psychological phenomenon that fits Crick's assumption and *probe* the corresponding neurobiological system to try to identify the neural differences between being aware and not being aware of the stimulus. This would give us a lead into the neural correlate of consciousness and hence into the mechanism. Fortunately, a property of the visual system known as *binocular rivalry* presents just the opportunity needed to proceed on Crick's assumption.[6]

What is binocular rivalry?

Suppose that you are looking at a computer monitor through special box with a division down the middle, so each eye sees only its half of the screen. If the two eyes are presented with the *same* stimulus, say a face, then what you see is one face. If, however, each eye gets different inputs—the left eye gets a face, and the right eye gets a sunburst pattern—then something quite surprising

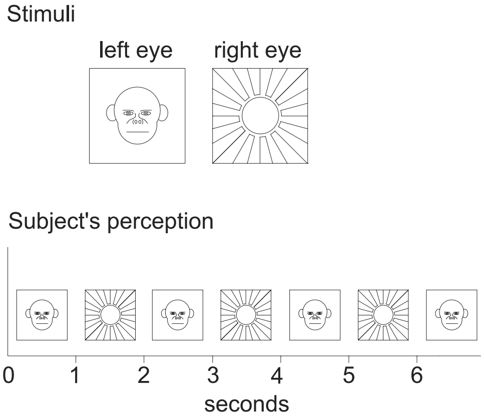

Figure 4.3 Bistable perception resulting from binocular rivalry. If different stimuli are presented to each eye, after a few moments of confusion, the brain settles down to perceiving the stimuli in an alternating sequence, where the perception of any given stimulus lasts only about 1 second. (Courtesy of P. M. Churchland.)

happens. After a few seconds, you perceive *alternating* stimuli: first sunburst, then face, then sunburst, then face. The perception is *bistable*, favoring neither one over the other, but switching back and forth between the two stimuli (figure 4.3). The reversal happens about once every 1–5 seconds, though the rate can be as long as once every 10 seconds. Many different stimuli give bistable perceptual effects, including horizontal bars shown to one eye and vertical bars to the other. So long as the stimuli are not too big or too small, the effect is striking, robust, and quite unambiguous.[7]

For the purposes of Crick's assumption, this setup is appealing: the opposing stimuli (e.g., the face and sunburst pattern) are *always* present, but the subject is *perceptually aware* of each only in alternating periods. Consider, for example, the face. It is *always* present, but now I am aware of the face, now I am aware of the sunburst pattern. Consequently, we can ask, What is the difference in the brain between those occasions when you *are* aware of the face and those when you are *not*?

Precisely *why* binocular rivalry exists is a question we leave aside for now, as there are various speculations but no definitive answer. It is fairly certain, however, that it is not a retinal or thalamic effect, but an effect of cortical processing. The most convincing hypothesis, favored by Leopold and Logothetis, is that binocular rivalry results from a system-level randomness that

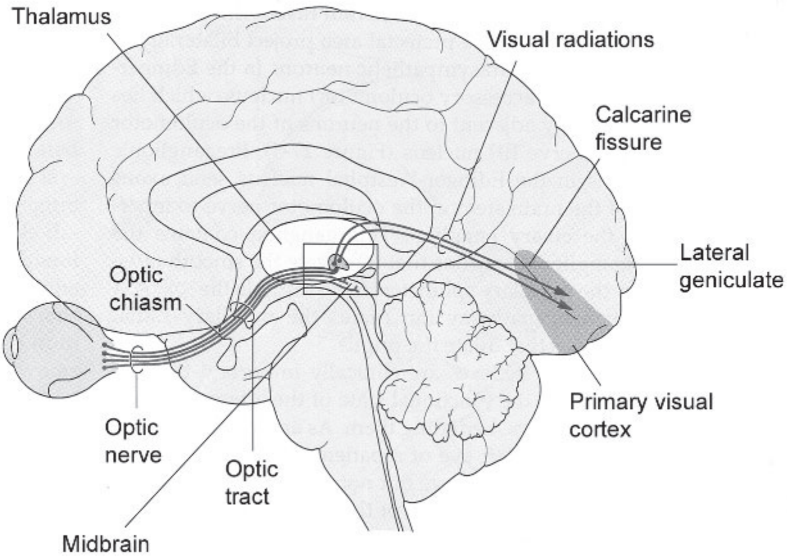

Figure 4.4 A diagram of human brain from the medial aspect showing the projections from the retina to the lateral geniculate nucleus of the thalamus and midbrain (superior colliculus and pretectum), and from the thalamus to cortical area V1 of the cerebral cortex. (Based on Kandel, Schwartz, and Jessell 2000.)

typifies exploratory behavior in general and whose function is to ensure that the brain does not get stuck in one perceptual hypothesis.[8]

On the neurobiological side, what is experimentally convenient about binocular rivalry is that in the visual system, cortical area STS (superior temporal sulcus) is known to contain individual neurons that respond preferentially to faces. This "tuning" of neurons, as it is called, is something that can be exploited by the experimentalist in the binocular rivalry setup (figures 4.4 to 4.6). This means that the cellular responses during presentation of rival stimuli can be recorded and monitored.

Area STS was identified, and its tuned neurons characterized, using single-neuron recording techniques in the monkey. This technique involves inserting a microelectrode into the cortex and recording the action potentials in the axon of a single neuron (figure 4.7).[9] On the basis of lesion data and fMRI studies, we know that human brains also have areas that are especially responsive to faces. Although such macrolevel data are extremely important, it has to be balanced by microlevel data from the single neuron. By and large, looking for single neurons whose activity correlates with conscious perception is something

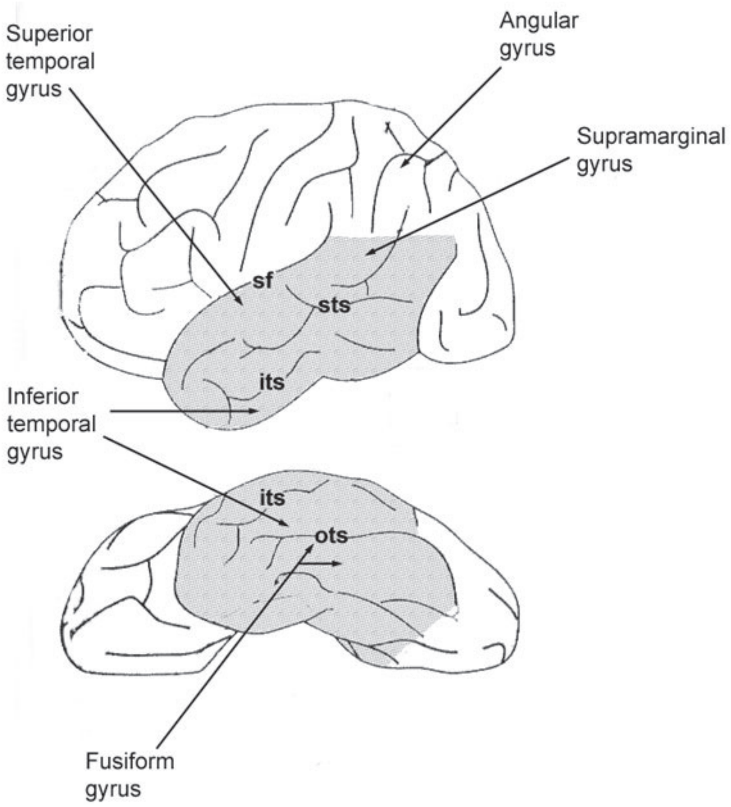

Figure 4.5 Schematic representations of the temporal lobe of human brain (shaded areas). The upper panel shows a side view (lateral aspect), and the lower panel shows the underside (ventral aspect). There are three general regions on the lateral surface of the temporal lobe: the superior temporal gyrus, the middle temporal gyrus, and the inferior temporal gyrus, which extends around to the ventral aspect of the temporal lobe. The ventral aspect includes the fusiform gyrus, also referred to as the occipitotemporal gyrus, and the parahippocampal gyrus, also referred to as the lingual gyrus. Abbreviations: its, inferior temporal sulcus; ots, occipitotemporal sulcus; sf, Sylvian fissure; sts, superior temporal sulcus. (Based on Rodman 1998.)

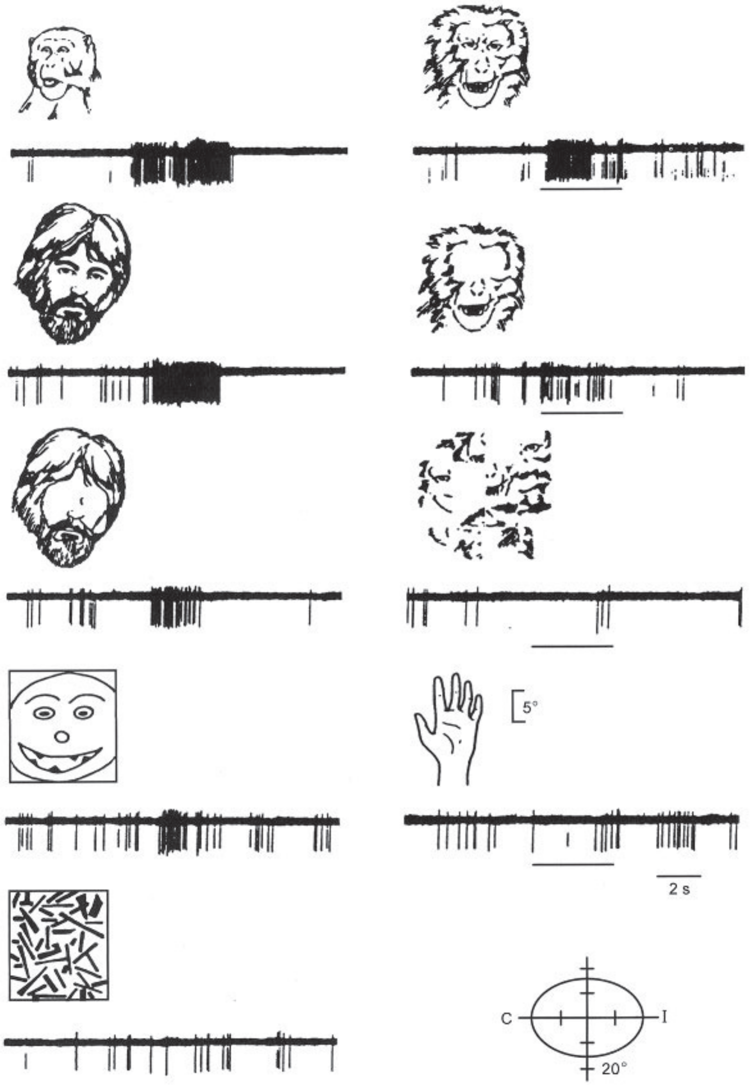

Figure 4.6 Recordings of activity of a cell with a large receptive field in the superior temporal gyrus as pictures of faces, degraded faces, or nonfaces are visually presented to a monkey. The cell responds most vigorously to faces, human or monkey or baboon. Activity is diminished if the eyes are removed or if the face's features are all present but jumbled. It responds better to a cartoon face than to the jumbled features or a nonface. When the monkey is shown a hand or a meaningless pattern, the cell response drops to its base firing rate. (From Bruce et al. 1981.)

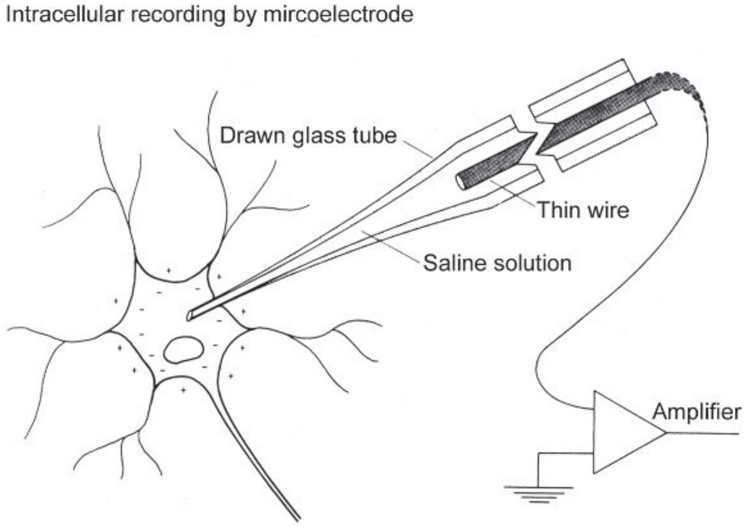

Figure 4.7 An idealized experiment for measuring the potential difference across a cell membrane. The electrode is a fine glass capillary with a tip no more than .1 micrometer in diameter, filled with a saline solution.

that must be done in monkeys. Nevertheless, by using an existing medical opportunity, Kreiman, Fried and Koch (2002) were able to repeat the Logothetis experiment in fourteen human surgical patients. Each had intractable epilepsy. To localize the seizure onset focus before surgery, eight depth electrodes were implanted in the medial temporal lobe of each patient. Recordings from these electrodes during bistable perception showed that about two thirds of the visually selective cells tracked the percept; none tracked the perceptually suppressed stimulus. Macaque monkeys are a good substitute for humans in the binocular-rivalry experiment because human and monkey brains are structurally very similar, and in particular, their visual systems are organizationally and structurally very similar. There is nevertheless a residual problem in using monkeys instead of humans, which is that humans can verbally answer "face" when they see a face, but the monkey cannot.

The tactic for overcoming this human/monkey difference is to train the monkey to respond by pressing a button with its left or right hand to indicate whether it sees a face or a sunburst. Monkeys are first trained in a standard (nonrivalrous) paradigm in which there is a correct answer and they are rewarded accordingly. That is the only way we have, so far, to let the monkey know what behavior we want. Once trained, monkeys are presented with rivalrous stimuli (face to one eye, sunburst to the other) to see how they respond. It is reassuring that monkeys' response behavior matches that of humans: it indicates an alternation in perception of the face versus the sunburst at about once per second.

A specific and significant doubt remains, nonetheless. Although monkeys may indeed be visually aware, they may not be using visual awareness to solve *this* problem. We know from human psychophysics that subjects can perform well above chance on a visual identification task even though they *report* that they are merely guessing their answers rather than judging on the basis of a conscious perception.

What adds fuel to this doubt is that the learning curves of the monkeys look like the learning curves of *operant conditioned* rats. In other words, we cannot assume that the experimenter's intent suddenly dawned on the monkey and it thought to itself, "Oh I get it. When I see *faces* I press *this* button, and when I see sunbursts I press *that* one!" and with that insight its performance jumps to nearly perfect. In fact, the monkeys show gradual improvement over days and even weeks rather than an abrupt improvement indicative of insight. The learning curves mean that the behavior of the animals is consistent with the *possibility* that connectivity is strengthened between visual area STS and motor cortex without visual awareness being part of the loop after all.

It is highly desirable to find ways to determine empirically, with a decent degree of probability, whether the animal uses *conscious* visual perception to solve the problem. Flexibility in response might be such an indicator. For example, if the monkey uses awareness to solve problems in anything like the way humans do, then the monkey should be able quickly to learn a new motor action to respond to the very same stimulus. If it uses both the new and the original response, the two should agree. The monkey should also appear surprised if a particular trial is easy and it gets the answer wrong. This sort of flexibility is characteristic of human conscious perception, and it is the kind of thing that should be demonstrable if the monkey is using visual awareness in solving the problem. Although we must shelve this problem for now, it is essential to acknowledge the need for developing experimental procedures on animals that overcome these problems.[10]

Inspired by the empirical problems confronting the experimentalist, the a priori skeptic might tender a much more tenacious skepticism about animal awareness. For example, the skeptic might complain that the monkey can only exhibit *behavior*, whereas the human can actually *talk*. So, the objection continues, we have no reason to think that the monkey is aware *at all*, *ever*, under *any* conditions.[11] The objection presupposes that speech is really a *direct* indication of consciousness, whereas button pressing is not.

Notice, first, that speech too is *just behavior*—behavior that humans have learned to perform. Even if the monkey did show verbal behavior, the determined skeptic would *still* complain that we could not be certain that its speech involves awareness as human speech does. Bonobo chimpanzees such as Kanzi and Pambanisha do display some verbal behavior, but the a priori skeptic waves this off as "mere conditioning."[12] We are now venturing into Skepticism, with a capital "S."

A thoroughly general Skepticism takes the form "How do I know that *any person*, let alone *some monkey*, is ever conscious? Indeed, how do I know that anything other than I exists? And moreover, how do I know that *I* was conscious before *this* very moment?" Part of the trouble with *this* brand of skepticism is that *no* empirical controls could allay the doubt one whit, *in principle*. The Skeptic thus overplays his hand, with the consequence that general Skepticism is hard to take very seriously beyond a moment or two.

A Skeptic can insist that there is no decisive proof that one is not dreaming, or that the universe was not created five minutes ago complete with fossil record, memories, history books, crumbling Roman ruins, and so on. Indeed, *there is no* decisive proof of the impossibility of what was just sketched. Still, as a hypothesis about reality, it is a bit silly.[13] *Specific* doubts about a *specific* experiment are a very different matter, however, and they do indeed have to be answered, one and all. In the absence of identifiable reasons for thinking that *only* humans can be visually aware, the similarities in monkey and human brains suggest that it is reasonable for me *provisionally* to assume that the monkey has visual awareness qualitatively *not very different* from ours. This is not a dogmatic declaration that monkeys are indeed visually aware as we are, but it is a useful *working assumption*, one that can sustain some interesting experiments. Nonetheless, it could be false, and it could be *falsified* empirically.

The binocular rivalry experiments

The neural correlates of visual awareness in binocular rivalry were first experimentally probed by neuroscientists Nikos Logothetis and Jeffrey Schall in 1989. Logothetis and Schall were using upward-moving and downward-moving gratings as stimuli. Their monkeys had been trained in advanced to indicate what they saw by pressing specific buttons, and the recording of single cells was done in visual cortical area MT. More recently (1997), Scheinberg and Logothetis have used a face and a sunburst pattern, and recorded in STS. Henceforth I shall frame the discussion around the face/sunburst stimuli, and I shall say "The monkey *sees* the face" as shorthand for "The monkey presses the button indicating its learned response to face stimuli," and so forth.

Simplified, the results are as follows. Consider a set of neurons, $N_1, \dots, N_5$, that were previously identified as responding preferentially to *faces*. (Suppose, for simplicity in this discussion, that faces are *always* present to the left eye, and sunbursts always to the right eye.) What do those neurons do when the monkey sees the *sunburst*? Some of them, perhaps N_1 and N_2, continue to respond, because of course the face is still present to the left eye, even if it is not consciously seen. Other face neurons, perhaps N_3 and N_4, do not respond. Now for the critical result: when and only when the monkey indicates that it *does see* the face, N_3 and N_4 respond (and as always, N_1, N_2 respond so long as the face is present) (figure 4.8).

Here is why this is interesting. Some neurons seem to be driven by the external stimulus; that is, they respond to the stimulus regardless of whether the monkey consciously perceives it. Others seem to respond only when the monkey sees—*consciously sees*—the stimulus. More exactly, the distribution of responses in STS was this: about 90 percent of the face neurons fire when and only when the monkey indicates it sees a face; the remainder always fire so long as the face is present on the monitor.

Can we say that the responsivity of the neurons in the 90 percent pool is *correlated with visual perception* (visual awareness)? Yes, but we need to go carefully here. Over a fairly generous time scale, "correlated with" could include events that are not identical with the state of perceptual awareness but are part of the causal sequence. More exactly, the data do not exclude the possibility that the responses of STS neurons are actually the causal antecedents—or possibly causal sequelae—of neural activity that *is* the awareness. In other words, we cannot simply conclude that this subset of STS neurons is the seat of visual awareness of faces. Progress has been made, but we do not want to overstate our conclusions.[14]

Although the binocular-rivalry experiments are a little complicated, they are important because they illustrate something that will surprise convention-bound philosophers. With the right experiment, you *can* make progress, even at the level of the single neuron, in investigating the neural causes or neural correlates of visual awareness. It shows, contra the naysayers, that headway, albeit only a little, is possible. Moreover, image data, using fMRI on humans, is consistent with the single-neuron results.[15] With further experiments, this beginning allows us to push on into territory that will be fruitful.

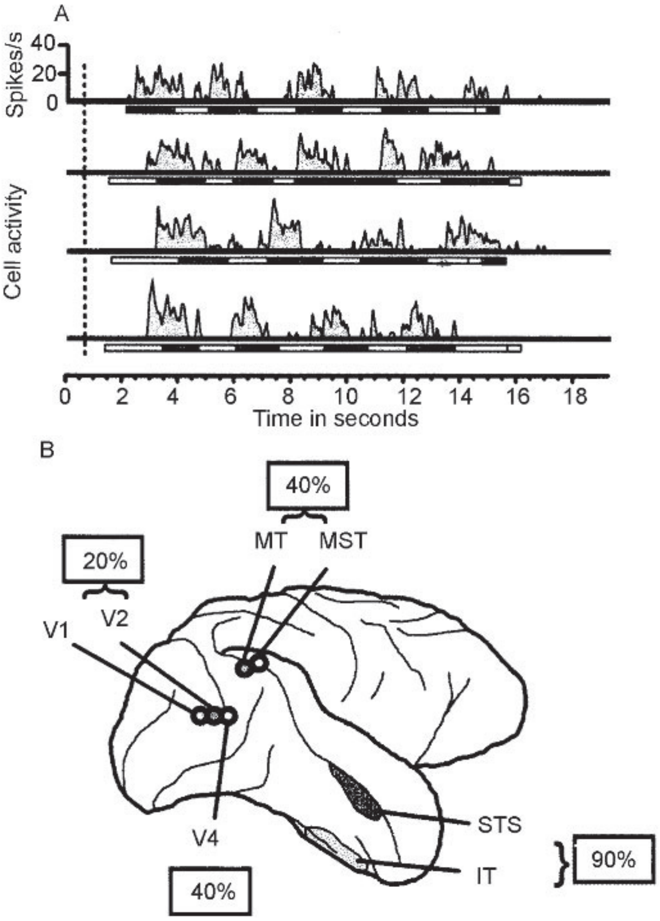

Figure 4.8 The neuronal responses of a face cell in the monkey brain during bistable perception. In the experiment, a monkey is trained to hold down one lever, e.g., the right-hand lever, when it sees a face, and to hold down the other lever when it sees a sunburst pattern. (A) The four horizontal graphs represent four observation periods, and the dashed vertical line indicates the onset of a rivalrous presentation (e.g., face and sunburst pattern). The animal's behavioral response is shown below the line, the shaded area representing the period during which animal holds down the appropriate lever. The cell response is shown above the line. The high rate of activity of the face cell begins just before, and ends just before, the period during which the animal holds down the face lever. The period of high activity (between 0 and 50 spikes/second) lasts for about 1 second. (B) The brain areas that contained the cells whose activity correlated with the monkey's subjective perception when responding to stimuli known to drive cells in that area. The greater the synaptic distance of the cortical area from the retina, the greater the percentage of cells driven by the subjective perception. Abbreviations: IT, inferior temporal; MT, middle temporal; MST, medial superior temporal sulcus; STS, superior temporal sulcus; V1, striate cortex; V2, V4, extrastriate cortex. (From Leopold and Logothetis 1999.)

Other experiments, similarly motivated, link up with the Logothetis results. Here is one strategy. To get a visual perception called "the waterfall illusion," you stare at a waterfall for several minutes. When you look away at a *still* surface, such as a gray blanket, you see upward motion, a kind of reverse, and illusory, waterfall. Roger Tootell used this phenomenon to run an experiment that complements the Logothetis and Schall experiment. The focus here will be on the neural correlates of conscious perception of upward motion induced in the *absence* of an externally present upward stimulus. Tootell used the noninvasive scanning technique fMRI to determine what cortical visual area showed greater activity when a human subject consciously perceives the waterfall illusion. He found, not unexpectedly, that motion-sensitive areas such as MT show increased activity with the onset of perception of the waterfall illusion. In this experiment too, it remains unknown whether MT neurons are actually neural correlates of consciousness, or whether they are just an element in the causal antecedents or consequences thereof.[16]

Hallucinations in human subjects present a different possibility for exploring what happens in the brain when a visual experience is present but the stimulus is *not*. Recently this has been elegantly pursued using fMRI by a group in London led by Ffytche.[17] Patients who suffer eye damage, for example as a result of detachment of the retina or glaucoma, lack normal vision. In some cases, these patients periodically experience highly vivid visual effects, though they are perfectly normal neuropsychiatrically. The character of the hallucinations varies from subject to subject, and unlike visual imagery, the visual objects appear to be in the outside world, and neither their appearance nor the nature of the visual image is under voluntary control.

One subject saw cartoonlike faces; another saw colored, shiny shapes rather like "futuristic cars." In the fMRI scanner, subjects signaled the onset of their visual hallucinations, and the scan data were analyzed. The data showed association of hallucinations with activity in the ventral visual regions, but with little activity in early visual cortex (V1). More specifically, if a subject hallucinated in color, an area independently identified as important in color processing was more active than if the hallucination was in black and white. Face hallucinations were associated with cortical subareas independently known to be involved with face processing, including the inferior temporal region (figure 4.9).

What do the Ffytche data mean? On their own, they do not solve the mystery, of course, but they are at least consistent with the data from binocular rivalry and from the waterfall illusion. These converging data suggest that a subset of neurons in visual cortical areas may support conscious visual perception.

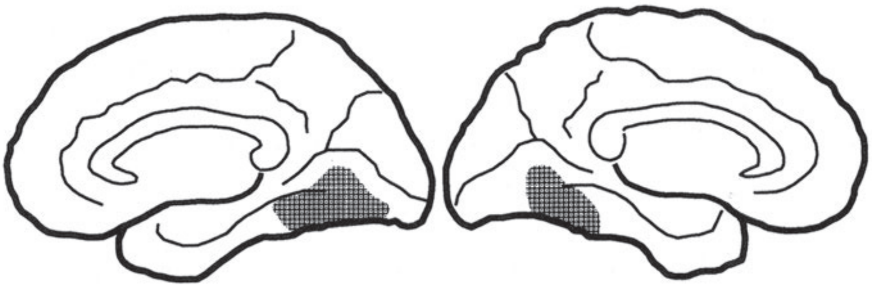

Figure 4.9 Bilateral lesions in the shaded region cause propopagnosia (loss of the capacity to identify individual faces). (Courtesy of Hanna Damasio.)

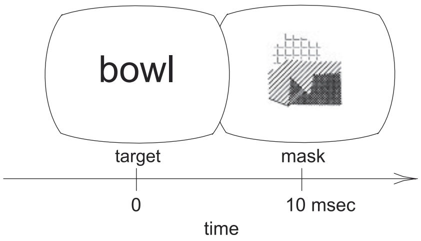

Figure 4.10 Visual masking. As the subject views the monitor, a word is presented, followed about 10 msec later by a noisy jumble—the mask. In these conditions, the subject sees only the mask, not the word.

Another experimental approach, also using fMRI, involves comparing brain activity during presentation of stimuli that are not consciously perceived and during presentation of stimuli that *are* consciously perceived. The experiments exploit an earlier behavioral result by Anthony Marcel, in which he showed that nonperceived stimuli had a quantifiable effect on subject's task performance. More specifically, Marcel flashed a word for about 10 msec., then immediately followed the word with a masking stimulus (a noisy visual stimulus flashed in the same location as the stimulus). The presentation of the mask somehow interferes with normal visual processing and the flashed item is not seen (figure 4.10). Subsequently, subjects were given a lexical-decision task, in which a string of letters was presented and the subject's task was to specify whether the string was or was not a word. Marcel showed that the subject's performance, measured in reaction time, was better for those words that had been presented in the *masked* condition than for words never presented. Moreover, processing of the flashed stimulus went beyond the mere physical shape of the stimulus because the effect was case-insensitive. (i.e., "BIRD" versus "bird"). This elegant experiment demonstrated a level of *semantic* processing even when subjects reported no conscious perception of the stimulus.

Dehaene and colleagues used the Marcel paradigm and recorded activity in normal subjects using fMRI in the masked and the visible conditions.[18] They showed that even in the masked condition, there is activity in both the fusiform gyrus and the precentral gyrus, areas that independent experiments indicate are active during conscious *reading* (see again fig. 4.9). In the condition where the stimulus was seen and not masked, the activity in the fusiform gyrus appeared to be about twelve times as strong as in the masked condition, and there was additional activity in the dorsolateral prefrontal cortex. The data suggest that the difference in brain activity in the two conditions is owed to conscious awareness of the stimuli.

Clever as the experiment is and important though the data are, several cautions are in order. First, the areas showing increased activity involve hundreds of millions of neurons, so the data are giving us a very general portrait, not detailed information about specific neurons or neuron-types and their role in awareness. Second, the data are consistent with the possibility that the greater activity in the nonmasked trial is caused by activation of a large range of neural networks whose stored information is associated with the flashed word. The mask may have associations too, but many fewer than a word. In the masked case, activation of networks associated with the word is probably interrupted by the mask, whereas the mask, being junk, provokes few associations. As the authors rightly note, the effects of the mask appear to start very early in the visual system, and propagate to higher levels. If the greater activity seen in the nonmasked case reflects greater numbers of activated associations, these associations might well be entirely nonconscious. They might be caused by a conscious representation, or by whatever it is that causes the conscious representation. Consequently, we cannot be sure that the greater range of activation in the unmasked case corresponds to conscious activity per se.[19]

Loops and conscious experience

An idea that has long been central to the approach of neuroscientist Gerald Edelman[20] is that loops (also referred to as *re-entrant pathways* and as *back projections*) are essential circuitry in the production of conscious awareness.[21] The idea is that some neurons carry signals from more peripheral to more central regions, such as from V1 to V2, while others convey more highly processed signals in the reverse direction, for example from V2 to V1. At an anatomical level, it is a general rule of cortical organization that forward-projecting neurons are matched by an equal or greater number of back-projecting neurons. Back-projecting neurons are a feature of brain organization generally, and in some instances, such as the pathway from V1 to the lateral geniculate nucleus (LGN) of the thalamus, back-projecting neurons are more numerous by a factor of ten than the forward-projecting neurons. Anatomically, then, the equipment is known to exist.

Why do Edelman and others think back projections have some particular role in consciousness? Part of the rationale for this point is that perception *always* involves classification; conscious seeing is *seeing as*.[22] Normally, one *sees* a fearful human face as fearful, rather than simply as a face followed by the explicit inference, "Aha, the eyes are especially wide open, etc., so this face is showing fear." In fact, most of us instantly recognize a fearful face but cannot articulate precisely what configuration of facial features is required for a face to show fear (figure 4.11). So we could not say what an *explicit inference* could use for *premises*, anyhow. Smells are often imbued with a *hedonic* dimension of meaning. The smell of rotten meat, for example, is disgusting to humans, whereas to vultures, it is appealing. Separating *in experience* the pure odor of rotten meat from the anhedonic nastiness of the smell is impossible.

Integrating hedonic components, emotional significance, associated cognitive representations, and so forth, with features of perception detected by the sensory systems almost certainly relies on loops—pathways projecting a signal back from structures such as the amygdala and hypothalamus (which have powerful roles in emotions and drives) to the sensory systems themselves, and pathways from so-called *higher areas* of cortex (e.g., prefrontal regions) to *lower areas* (e.g., V1). That we directly perceive a face with its fearful expression implies that information about the emotion must be routed back to the visual system at some level. A purely feedforward neural network cannot achieve this kind of integration.

Artificial neural network (ANN) research indicates that many of the consciousness-related functions—STM, attention, sensory perception, meaning—are handled most powerfully and efficiently by networks with recurrent projections. The range of functions that back projections perform has not been precisely demonstrated in real neural networks, and there are serious technical difficulties to be overcome before back-projection physiology can get very far. Nevertheless, the fact that back projections in ANNs render those systems vastly more powerful, and more powerful *in the ways relevant to consciousness-related functions*, is highly suggestive.[23]

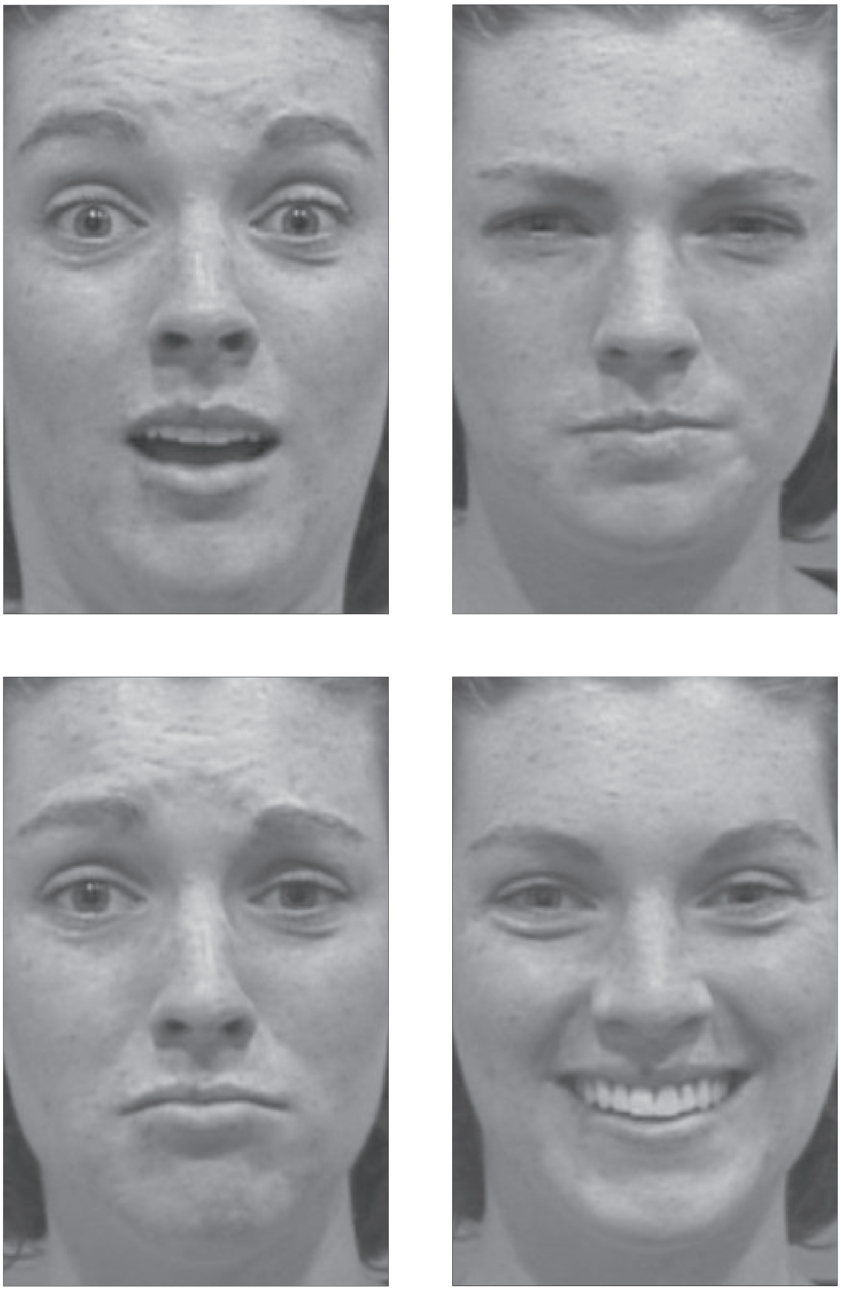

Figure 4.11 Human facial expressions of four emotions: fear, anger, sadness, and happiness. (Faces courtesy of Dailey, Cottrell, and Reilly. Copyright 2001 California Facial Expressions Database [CAFE].)

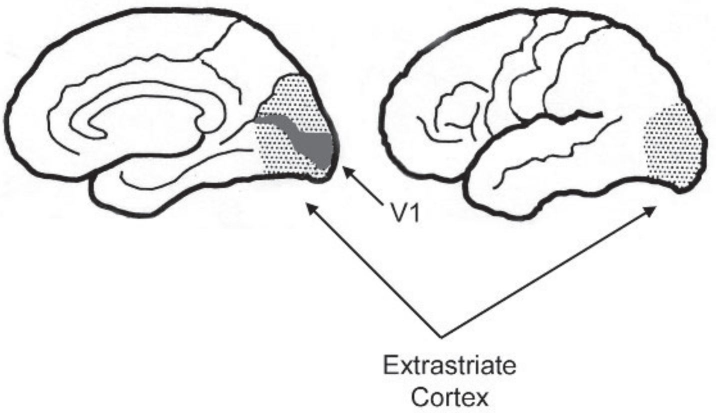

Figure 4.12 A schematic diagram of the human brain showing the position of V1. On the left is the medial view; on the right is the lateral view. In the visual cortex, V1 is located in the calcarine sulcus in the medial aspect, shown in dark shading. The extra striate cortex is shown with dotted shading. (Courtesy of Hanna Damasio.)

Experimental evidence is beginning to come in to support this idea. For example, Pascual-Leone and Walsh exploited the fact that transcranial magnetic stimulation (TMS) of cortical visual area V1 will cause the subject to experience small flashes of light, while stimulation of cortical visual area MT will produce flashes of light that move.[24] The anatomical fact of importance is that there are back projections from MT to V1. (In fact the back projections typical of cortical organization are also seen in the brainstem and spinal cord, as well as in structures such as the hypothalamus. They are essentially everywhere.) So here is their experiment: stimulate MT in a manner normally adequate to produce moving flashes of light, and also stimulate V1, but at an intensity so low that it does not cause perception of lights, but high enough to interfere with the normal effect of back-projected signals from MT. If back-projected signals from MT are necessary to see moving flashes, then in this condition, no moving flashes will be seen. These are indeed the results. Subjects see flashes, but not moving flashes.

As always, optimism must be tempered with skeptical questions. One major question concerns what exactly is the effect of TMS at the neuronal level, how focal the stimulation really is, and how far the effect spreads, cortically and subcortically. A further problem arises from the nature of human brain anatomy. In the macaque monkey, V1 is on the dorsal surface of the brain. In human brains, V1 is on the *medial* surface of the occipital lobe (figure 4.12).

Consequently, if you aim to stimulate V1 with TMS, you will also stimulate the dorsal regions, and activity in the pathways from the incidentally stimulated areas can be predicted to affect both V1 and V2. The worry is that the incidentally stimulated areas confound the results.

In any case, even if back projections are *necessary* for consciousness, we know that they are not *sufficient*. Back projections function in phylogenetically older parts of the brain, such as the spinal cord; some are active when subjects are under anesthesia, in a deep sleep, or in a coma. If a subset of cortical back projections are indeed subserving awareness of visual stimuli, it will be important to determine which axons they are and what precisely is the nature of their signals.

Theorizing and narrowing the hypothesis space

In addition to designing experiments to identify the neural correlates of consciousness, pulling together data bearing on the conditions for visual experience and isolating structural and functional constraints can help narrow the hypothesis space. Especially in the early stages of the problem, this is a very useful strategy, particularly because some of the concepts needed to articulate a good hypothesis undoubtedly need to be invented as the search space narrows ever more.

Loops are likely to be one structural constraint on the substrate for consciousness. As Francis Crick and Christof Koch suggest, other constraints that emerge from the experimental literature include the following:[25]

- The neurons whose collective activity constitutes being aware of something are distributed spatially. Transiently, they form a "coalition" that lasts for the duration of the awareness of a particular perception, such as visual awareness of Lincoln's face. Individual neurons can be elements in different coalitions as a function of the percepts. For example, a particular neuron might be part of a coalition that constitutes being aware of Lincoln's face, but it also might be part of a coalition that constitutes being aware of a human hand, or a coalition for a dog face.
- Neurons in the coalition whose activity constitutes a perceptual awareness probably need to reach a threshold in order for the coalition's activity to constitute perceptual awareness.
- Normally, though perhaps not necessarily, a coalition emerges as a consequence of synchrony of firing in neuron populations that project to the

coalition members. This synchrony of firing is part of the causal conditions for reaching the threshold.

- When neurons involved in perceptual awareness do fire above that threshold, they continue firing for a short but sustained period of time (e.g., longer than 100 milliseconds but not as long as a minute).
- Attention probably up-regulates the activity of the relevant neurons, getting them closer to their threshold.
- In awareness of a certain visual phenomenon, say the face of Lincoln, some neurons will be activated as part of the cognitive background, while some will be activated as essential to the experience itself. These latter neurons Crick and Koch call "essential nodes," to distinguish them from neurons that contribute to the cognitive background. Included in the cognitive background are the *expectation* that the face is the front of the head, and various nonconscious, *tacit beliefs*, e.g., that if Lincoln had been born in Australia, he would not have been president of the United States. The cognitive background includes also various *associations* and *inferential* connections, for example, the association with the civil war, and the capacity to infer from "Lincoln was president of the United States in 1864," the statement that "Lincoln is not now president of the United States."
- At any given moment there is probably a competition between various essential-node neurons for which neurons will fire at the threshold and hence which representation will be conscious. Thus, if I am paying close attention to events on television, I may not hear the lawnmower running next door. This implies that the essential-node neurons in the auditory system will have lost out in the competition to those in the visual system representing the events on the television.

Ideally, the items in this list will jell to form a kind of prototheory of neural mechanisms supporting perceptual awareness. In the role of prototheory, the list may provoke experiments to confirm or disconfirm any one of its items, and thus move us closer to understanding the nature of consciousness. Having some sort of theoretical scaffolding is a clear improvement over groping haphazardly. Even if none of the items on the list turns out to be part of the explanation of consciousness, the exercise is valuable, because it orients us toward thinking of the problem of consciousness in terms of *mechanisms*, that is, in terms of causal organization. Identifying neural correlates is one thing, and likely a useful thing, but the goal we ultimately want to reach is identifying causal mechanisms so as to understand *how* consciousness occurs.

A methodological question about neural correlates

In the foregoing experiments, there was evidence of neural activity correlated with conscious awareness. Nevertheless, I expressed caution concerning what such correlational evidence signifies. The major reason has already been stated: finding correlations between neural activity and a subject's reports of perceptual awareness is *consistent* with any of the following: (1) the neural activity is a background condition for perceptual awareness, (2) the neural activity is part of the cause, (3) the neural activity is part of the sequelae of the awareness, (4) the neural activity parallels, but plays no direct role in, perceptual awareness, and (5) the neural activity is what perceptual awareness can be *identified* with (the *identificand*).

Ultimately, if we want to be able to explain the nature of consciousness in neural terms, what we seek is the *identification* of some class of neural activity with perceptual awareness. That is, we want our data to justify interpretation (5). As is evident, however, correlational data per se do not rule out all alternatives except (5). That some event *x* is a correlate of some phenomenon *y* does tell you a *little*, such as that you *may* be on the right path for finding the *identificand*. For similar reasons, that some event *z* fails to correlate with some phenomenon *y* suggests that you may be on the wrong path. This is not the whole pudding, nor is it nothing, and one has to start somewhere.

Determining that two phenomena are systematically correlated requires testing under a wide range of conditions. It is not enough, for example, to get fMRI data showing that in awake subjects, a specific cortical visual area is highly active whenever the subject reports visual awareness of an object. We want also to know whether there is activity in that brain region when the subject is not conscious. For example, it is essential to know whether the brain of a subject in a coma or in a persistent vegetative state or under anesthesia shows activity in that brain region when a visual stimulus is presented. This is not idle skepticism. Activity in various cortical areas is known to occur in response to an external stimulus in precisely these unusual conditions. A patient in a persistent vegetative state, for example, exhibits no signs of awareness, and in particular, no behavioral sign of awareness when shown a familiar person. Nevertheless, when the subject was shown familiar faces, the so-called "face area" of the cortex showed a pattern of increased activity similar to that of the normal subject.[26] As Damasio correctly notes, such data are powerful clues that neurons in the visual cortex may not be the generators of visual conscious experience. Rather, their activities are representations that the subject might be aware of *if* he were conscious. So until the tough cases have been excluded by experiment, no conclusion can be drawn from correlations in the relatively easy cases.

There is, however, the deeper problem touched on earlier; it is the problem of knowing what you are looking at. It is reasonable to hope that there is a class of neural activity correlated *always* and *only* with perceptual awareness, and that such activity is *identifiable* as conscious awareness. Nonetheless, even if there is such a class of activity, knowing that *this* measured activity belongs to *that class* may be discoverable only very indirectly. In other words, we might be looking straight at an instance of the class without in the slightest recognizing that it is an instance. This will happen if, as is very likely, the physical substrate does not have a property that is salient to the naive observer, but is recognizable only through the lens of a more comprehensive *theory* of brain function.

An analogy may make this point clearer. In the nineteenth century, the nature of light was a profound mystery. Suppose, to be fanciful, that nineteenth-century physicists address the mystery by seeking the microstructural correlates of light. They hope that there is a particular class of microstructural phenomena that is *always* and *only* correlated with light, and that such activity, or something connected to it, is *identifiable* as light. The rough idea is to look for the "defining property"—the *identificand*, as we may refer to it.

Since those of us living now have the benefit of post-Maxwellian physics, we know that the defining property is characterized abstractly and nonobservationally by the theory of electromagnetic radiation. That is, Maxwell realized that the equations characterizing light matched perfectly the equations characterizing radio waves, x-rays, and other electromagnetic phenomena. He rightly concluded that light just *is* yet another form of electromagnetic radiation. Observable properties give no hint of this, but the match of deep, unobservable properties gave the game away.

Here is the question: could our imagined pre-Maxwellian correlation hunters notice, even if they looked closely, that radio waves and light share that same deep property? Probably not, since, until they understand a good deal more about electromagnetic radiation, they lack the conceptual resources to see what *counts as the same property*. This is because they do not yet have the slightest inkling that light *is* electromagnetic radiation, or that x-rays, gamma rays, etc., even exist (see plate 1).

Or think of the problem this way: How would *you* know, independently of Lavosier's work on oxygen, that rusting, metabolizing, and burning are the same microphysical process, but that sunlight and lightning are *not*? What property would you look at? And if you did by luck make a guess that the first three phenomena share a microstructural property, how would you test your idea?

This is *not* to say that looking for the neural correlates of consciousness is futile. On the contrary, at this very early stage of the neurobiological investigation of consciousness, it is undoubtedly wise to give it the best shot possible. My point is that it is also wise to recognize the pitfalls and to appreciate that they are not merely technological, but derive also from the absence of a firmly planted *theoretical* framework for understanding how the brain works.[27]

The experiments discussed in this section, and others with a similar general conceptual slant, are important because they have opened doors. From the vantage point of 1980, when such experiments were barely conceivable, they look downright spectacular. At the very least, they inspire researchers to invent better and better experimental designs. It should be noted, however, that the examples in this section do share a certain conceptual slant that is open to criticism. All are focused mainly on the *cerebral cortex*, and all are drawn from the *visual system*. This narrowing of the focus can be valuable, especially when different experimental strategies unearth complementary results, as those discussed above do to some extent. Focusing narrowly allows us to probe deeply, if not broadly, and that can be rewarding.

Nevertheless, for all we can tell now, it could turn out that other modalities play a role in consciousness that is more straightforward and less complicated than the role of vision. Possibly, exploration of olfactory or somatosensory processing will reveal principles obscured thus far. More seriously, it could turn out that it is not *cortical* neurons—or not cortical neurons *alone*—whose activity is identifiable with awareness, but rather, the activity of various *non-cortical* neurons in the brainstem, thalamus, hypothalamus, and so forth.[28] It is common knowledge that subcortical activity does figure in the causal antecedents. Whether some subcortical activity is more than that, however, is a possibility we shall explore in section 1.4.

1.4 The Indirect Approach

Attention, short-term memory, autobiographical memory, self-representation, perception, imagery, thought, meaning, being awake, self-referencing—*all*