12

QUALIA DISQUALIFIED

1. A NEW KITE STRING

Thrown into a causal gap, a quale will simply fall through it.

IVAN FOX (1989), p. 82

When your kite string gets snarled up, in principle it can be unsnarled, especially if you're patient and analytic. But there's a point beyond which principle lapses and practicality triumphs. Some snarls should just be abandoned. Go get a new kite string. It's actually cheaper in the end than the labor it would take to salvage the old one, and you get your kite airborne again sooner. That's how it is, in my opinion, with the philosophical topic of qualia, a tormented snarl of increasingly convoluted and bizarre thought experiments, jargon, in-jokes, allusions to putative refutations, "received" results that should be returned to sender, and a bounty of other sidetrackers and time-wasters. Some messes are best walked away from, so I am not going to conduct an analytical tour of that literature, even though it contains moments of insight and ingenuity from which I have benefited (Shoemaker, 1975, 1981, 1988; White, 1986; Kitcher, 1979; Harman, 1990; Fox, 1989). I've tried in the past to unsnarl the issue (Dennett, 1988a), but now I think it's better if we try to start over almost from scratch.

It's not hard to see how philosophers have tied themselves into such knots over qualia. They started where anyone with any sense

would start: with their strongest and clearest intuitions about their own minds. Those intuitions, alas, form a mutually self-supporting closed circle of doctrines, imprisoning their imaginations in the Cartesian Theater. Even though philosophers have discovered the paradoxes inherent in this closed circle of ideas — that's why the literature on qualia exists — they haven't had a *whole alternative vision* to leap to, and so, trusting their still-strong intuitions, they get dragged back into the paradoxical prison. That's why the literature on qualia gets more and more convoluted, instead of resolving itself in agreement. But now we've put in place just such an alternative vision, the Multiple Drafts model. Using it, we can offer a rather different positive account of the issues. Then we can pause in sections 4 and 5 to compare it to the visions I hope it will replace.

An excellent introductory book on the brain contains the following passage:

> "Color" as such does not exist in the world; it exists only in the eye and brain of the beholder. Objects reflect many different wavelengths of light, but these light waves themselves have no color. [Ornstein and Thompson, 1984, p. 55]

This is a good stab at expressing the common wisdom, but notice that taken strictly and literally, it cannot be what the authors mean, and it cannot be true. Color, they say, does not exist "in the world" but only in the "eye and brain" of the beholder. But the eye and brain of the beholder are in the world, just as much parts of the physical world as the objects seen by the observer. And like those objects, the eye and brain are colorful. Eyes can be blue or brown or green, and even the brain is made not just of gray (and white) matter: in addition to the *substantia nigra* (the black stuff) there is the *locus ceruleus* (the blue place). But of course the colors that are "in the eye and brain of the beholder" in *this* sense are not what the authors are talking about. What makes anyone think there is color in any other sense?

Modern science — so goes the standard story — has removed the color from the physical world, replacing it with colorless electromagnetic radiation of various wavelengths, bouncing off surfaces that variably reflect and absorb that radiation. It may look as if the color is out there, but it isn't. It's *in here* — in the "eye and brain of the beholder." (If the authors of the passage were not such good materialists, they would probably have said that it was in the *mind* of the observer, saving themselves from the silly reading we just dismissed, but creating even worse problems for themselves.) But now, if there is no inner *figment*

that could be colored in some special, subjective, in-the-mind, phenomenal sense, colors seem to disappear altogether! Something has to be the colors we know and love, the colors we mix and match. Where oh where can they be?

This is the ancient philosophical conundrum we must now face. In the seventeenth century, the philosopher John Locke (and before him, the scientist Robert Boyle) called such properties as colors, aromas, tastes, and sounds *secondary qualities*. These were distinguished from the primary qualities: size, shape, motion, number, and solidity. Secondary qualities were not themselves things-in-the-mind but rather the powers *of* things in the world (thanks to their particular primary qualities) to produce or provoke certain things in the minds of normal observers. (And what if there were no observers around? This is the eternally popular puzzler about the tree in the forest that falls. Does it make a sound? The answer is left as an exercise for the reader.) Locke's way of defining secondary qualities has become part of the standard layperson's interpretation of science, and it has its virtues, but it also gives hostages: the things produced in the mind. The secondary quality *red*, for instance, was for Locke the dispositional property or power of certain surfaces of physical objects, thanks to their microscopic textural features, to produce in us the *idea of* red whenever light was reflected off those surfaces into our eyes. The power in the external object is clear enough, it seems, but what kind of a thing is an idea of red? Is it, like a beautiful gown of blue, *colored* — in some sense? Or is it, like a beautiful discussion of purple, just *about* a color, without itself being colored at all? This opens up possibilities, but how could an idea be just *about* a color (e.g., the color red) if nothing anywhere is red?

What is red, anyway? What are colors? Color has always been the philosophers' favorite example, and I will go along with tradition for the time being. The main problem with the tradition nicely emerges in the philosophical analysis of Wilfrid Sellars (1963, 1981b), who distinguished the dispositional properties of objects (Locke's secondary qualities) from what he called occurrent properties. A pink ice cube in the freezer with the light off has the secondary quality pink, but there is no instance of the property occurrent pink until an observer opens the door and looks. Is occurrent pink a property of something in the brain or something "in the external world"? In either case, Sellars insisted, occurrent pink is a "homogeneous" property of something real. Part of what he meant to deny by this insistence on homogeneity would be the hypothesis that occurrent pink is anything like neural *activity of* intensity *97* in region *75* of the brain. He also meant to deny

that the subjective world of color phenomenology is exhausted by anything as colorless as judgments that one thing or another is, or seems to be, pink. For instance, the act of recalling in your mind's eye the color of a ripe banana and judging that it is the color yellow would not by itself bring into existence an instance of occurrent yellow (Sellars, 1981; Dennett, 1981b). That would merely be judging that something was yellow, a phenomenon that by itself is as devoid of occurrent yellow as a poem about bananas would be.

Sellars went so far as to claim that all of the physical sciences would have to be revolutionized to make room for occurrent pink and its kin. Few philosophers went along with him on this radical view, but a version of it has recently been resurrected by the philosopher Michael Lockwood (1989). Other philosophers, such as Thomas Nagel, have supposed that even revolutionized science would be unable to deal with such properties:

> The subjective features of conscious mental processes — as opposed to their physical causes and effects — cannot be captured by the purified form of thought suitable for dealing with the physical world that underlies the appearances. [1986, p. 15]

Philosophers have adopted various names for the things in the beholder (or properties of the beholder) that have been supposed to provide a safe home for the colors and the rest of the properties that have been banished from the "external" world by the triumphs of physics: "raw feels," "sensa," "phenomenal qualities," "intrinsic properties of conscious experiences," "the qualitative content of mental states," and, of course, "qualia," the term I will use. There are subtle differences in how these terms have been defined, but I'm going to ride roughshod over them. In the previous chapter I seemed to be denying that there are any such properties, and for once what seems so is so. I am denying that there are any such properties. But (here comes that theme again) I agree wholeheartedly that there seem to be qualia.

There seem to be qualia, because it really does seem as if science has shown us that the colors can't be out there, and hence must be in here. Moreover, it seems that what is in here can't just be the judgments we make when things seem colored to us. This reasoning is confused, however. What science has actually shown us is just that the light-reflecting properties of objects cause creatures to go into various discriminative states, scattered about in their brains, and underlying a host of innate dispositions and learned habits of varying complexity. And what are *their* properties? Here we can play Locke's card a second time:

These discriminative states of observers' brains have various "primary" properties (their mechanistic properties due to their connections, the excitation states of their elements, etc.), and in virtue of these primary properties, they have various secondary, merely dispositional properties. In human creatures with language, for instance, these discriminative states often eventually dispose the creatures to express verbal judgments alluding to the "color" of various things. When someone says "I know the ring isn't really pink, but it sure seems pink," the first clause expresses a judgment about something in the world, and the second clause expresses a second-order judgment about a discriminative state about something in the world. The semantics of such statements makes it clear what colors supposedly are: reflective properties of the surfaces of objects, or of transparent volumes (the pink ice cube, the shaft of limelight). And that is just what they are in fact — though saying just *which* reflective properties they are is tricky (for reasons we will explore in the next section).

Don't our internal discriminative states *also* have some special "intrinsic" properties, the subjective, private, ineffable, properties that constitute *the way things look* to us (sound to us, smell to us, etc.)? Those additional properties would be the qualia, and before looking at the arguments philosophers have devised in an attempt to prove that there are these additional properties, we will try to remove the motivation for believing in these properties in the first place, by finding alternative explanations for the phenomena that seem to demand them. Then the systematic flaws in the attempted proofs will be readily visible.

According to this alternative view, colors are properties "out there" after all. In place of Locke's "ideas of red" we have (in normal human beings) discriminative states that have the content: *red*. An example will help make absolutely clear what these discriminative states are — and more important, what they are not. We can compare the colors of things in the world by putting them side by side and looking at them, to see what judgment we reach, but we can also compare the colors of things by just recalling or imagining them "in our minds." Is the standard red of the stripes on the American flag the same red as, or is it darker or lighter or brighter or more or less orange than, the standard red of Santa Claus's suit (or a British pillar box or the Soviet red star)? (If no two of these standards are available in your memory, try a different pair, such as Visa-card blue and sky blue, or billiard-table-felt green and Granny-Smith-apple green, or lemon yellow and butter yellow.) We are able to make such comparisons "in our

mind's eyes," and when we do, we somehow make something happen in us that retrieves information from memory and permits us to compare, in conscious experience, the colors of the standard objects as we remember them (as we take ourselves to remember them, in any case). Some of us are better at this than others, no doubt, and many of us are not very confident in the judgments we reach under such circumstances. That is why we take home paint samples, or take fabric samples to the paint store, so that we can put side by side in the external world instances of the two colors we wish to compare.

When we do make these comparisons "in our mind's eyes," what happens, according to my view? Something strictly analogous to what would happen in a machine — a robot — that could also make such comparisons. Recall from chapter 10 the CADBLIND Mark I Vorsetzer (the one with the camera that could be aimed at the CAD screen). Suppose we put a color picture of Santa Claus in front of it and ask it whether the red in the picture is deeper than the red of the American flag (something it has already stored in its memory). This is what it would do: retrieve its representation of Old Glory from memory, and locate the "red" stripes (they are labeled "red #163" in its diagram). It would then compare this red to the red of the Santa Claus suit in the picture in front of its camera, which happens to be transduced by its color graphics system as red #172. It would compare the two reds by *subtracting 163 from 172 and getting 9*, which it would interpret, let's say, as showing that Santa Claus red seems somewhat deeper and richer (to it) than American flag red.

This story is deliberately oversimple, to dramatize the assertion I wish to make: It is obvious that the CADBLIND Mark I doesn't use figment to render its memory (or its current perception), *but neither* do *we*. The CADBLIND Mark I probably doesn't know how it compares the colors of something seen with something remembered *and neither* do *we*. The CADBLIND Mark I has — I will allow — a rather simple, impoverished color space with few of the associations or built-in biases of a human being's personal color space, but aside from this vast difference in dispositional complexity, there is no important difference. I could even put it this way: There is no *qualitative* difference between the CADBLIND's performance of such a task and our own. The discriminative states of the CADBLIND Mark I have content in just the same way, and for just the same reasons, as the discriminative brain states I have put in place of Locke's ideas. The CADBLIND Mark I *certainly* doesn't have any qualia (at least, that is the way I expect lovers of qualia to jump at this point), so it does indeed follow from

my comparison that I am claiming that we don't have qualia either. The sort of difference that people imagine there to be between any machine and any human experiencer (recall the wine-tasting machine we imagined in chapter 2) is one I am firmly denying: There is no such sort of difference. There just seems to be.

2. WHY ARE THERE COLORS?

When Otto, in chapter 11, judged that there seemed to be a glowing pinkish ring, what was the content of his judgment? If, as I have insisted, his judgment wasn't about a quale, a property of a "phenomenal" seeming-ring (made out of figment), just what was it about? What property did he find himself tempted to attribute (falsely) to something out in the world?

Many have noticed that it is curiously difficult to say just what properties of things in the world colors could be. The simple and appealing idea — still found in many elementary discussions — is that each color can be associated with a unique wavelength of light, and hence that the property of being red is simply the property of reflecting all the red-wavelength light and absorbing all the other wavelengths. But this has been known for quite some time to be false. Surfaces with different fundamental reflective properties can be seen as the same color, and the same surface under different conditions of lighting can be seen as different colors. The wavelengths of the light entering the eye are only indirectly related to the colors we see objects to be. (See Gouras, 1984; Hilbert, 1987; and Hardin, 1988, for surveys of the details with different emphases.) For those who had hoped there would be some simple, elegant way to cash in Locke's promissory note about dispositional powers of surfaces, the situation could hardly be more bleak. Some (e.g., Hilbert, 1987) have decided to anchor color objectively by declaring it to be a relatively straightforward property of external objects, such as the property of "surface spectral reflectance"; having made that choice, they must then go on to conclude that normal color vision often presents us with illusions, since the constancies we perceive match up so poorly with the constancies of surface spectral reflectance measured by scientific instruments. Others have concluded that color properties are best considered subjectively, as properties to be defined strictly in terms of systems of brain states in observers, ignoring the confusing variation in the world that gives rise to these states: "Colored objects are illusions, but not unfounded illusions. We are normally in chromatic perceptual states, and these are neural states"

(Hardin, 1988, p. 111; see Thompson, Palacios, and Varela, in press, for a critical discussion of these options, and further arguments for the better option to be adopted here).

What is beyond dispute is that there is no simple, nondisjunctive property of surfaces such that all and only the surfaces with that property are red (in Locke's secondary quality sense). This is an initially puzzling, even depressing fact, since it seems to suggest that our perceptual grip on the world is much worse than we had thought — that we are living in something of a dream world, or are victims of mass delusion. Our color vision does not give us access to simple properties of objects, even though it seems to do so. Why should this be so?

Just bad luck? Second-rate design? Not at all. There is a different, and much more illuminating, perspective we can take on color, first shown to me by the philosopher of neuroscience, Kathleen Akins (1989, 1990).[1] Sometimes new properties come into existence for a reason. A particularly useful example is provided by the famous case of Julius and Ethel Rosenberg, who were convicted and executed in 1953 for spying on the U.S. atomic bomb project for the Soviet Union. It came out at their trial that at one point they improvised a clever password system: a cardboard Jell-O box was torn in two, and the pieces were taken to two individuals who had to be very careful about identifying each other. Each ragged piece became a practically foolproof and unique "detector" of its mate: at a later encounter each party could produce his piece, and if the pieces lined up perfectly, all would be well. Why does this system work? Because tearing the cardboard in two produces an edge of such informational complexity that it would be virtually impossible to reproduce by deliberate construction. (Note that cutting the Jell-O box with straight-edge and razor would entirely defeat the purpose.) The particular jagged edge of one piece becomes a *practically* unique pattern-recognition device for its mate; it is an apparatus or transducer for detecting the shape property *M*, where *M* is uniquely instantiated by its mate.

In other words, the shape property *M* and the *M*-property-detector that detects it were made for each other. There would be no reason for either to exist, to have been created, in the absence of the other. And the same thing is true of colors and color vision: they were made for each other. Color-coding is a fairly recent idea in "human factors engineering," but its virtues are now widely recognized. Hospitals lay out

1. Variations on these themes can be found in Humphrey (1976, 1983a) and in Thompson, Palacios, and Varela (in press).

colored lines in the corridors, simplifying the directions that patients must follow: "To get to physiotherapy, just follow the yellow line; to get to the blood bank, follow the red line!" Manufacturers of televisions, computers, and other electronic gear color-code the large bundles of wires inside so that they can be easily traced from point to point. These are recent applications, but of course the idea is much older; older than the Scarlet Letter with which an adulterer might be marked, older than the colored uniforms used to tell friend from foe in the heat of battle, older than the human species, in fact.

We tend to think of color-coding as the clever introduction of "conventional" color schemes designed to take advantage of "natural" color vision, but this misses the fact that "natural" color vision coevolved *from the outset* with colors whose *raison d'être* was color-coding (Humphrey, 1976). Some things in nature "needed to be seen" and others needed to see them, so a system evolved that tended to minimize the task for the latter by heightening the salience of the former. Consider the insects. Their color vision coevolved with the colors of the plants they pollinated, a good trick of design that benefited both. Without the color-coding of the flowers, the color vision of the insects would not have evolved, and vice versa. So the principle of color-coding is the basis of color vision in insects, not just a recent invention of one clever species of mammal. Similar stories can be told about the evolution of color vision in other species. While some sort of color vision may have evolved initially for the task of discriminating inorganic phenomena visually, it is not yet clear that this has happened with any species on this planet. (Evan Thompson has pointed out to me that honeybees may use their special brand of color vision in navigation, to discriminate polarized sunlight on cloudy days, but is this a secondary utilization of color vision that originally coevolved with flower colors?)

Different systems of color vision have evolved independently, sometimes with radically different color spaces. (For a brief survey, and references, see Thompson, Palacios, and Varela, in press.) Not all creatures with eyes have any sort of color vision. Birds and fish and reptiles and insects clearly have color vision, rather like our "trichromatic" (red-green-blue) system; dogs and cats do not. Among mammals, only primates have color vision, and there are striking differences among them. Which species have color vision, and why? This turns out to be a fascinating and complex story, still largely speculative.

Why do apples turn red when they ripen? It is natural to assume that the entire answer can be given in terms of the chemical changes

that happen when sugar and other substances reach various concentrations in the maturing fruit, causing various reactions, and so forth. But this ignores the fact that there wouldn't be apples in the first place if there weren't apple-eating seed-spreaders to see them, so the fact that apples are readily visible to at least some varieties of apple-eaters is a condition of their existence, not a mere "hazard" (from the apple's point of view!). The fact that apples have the surface spectral reflectance properties they do is as much a function of the photopigments that were available to be harnessed in the cone cells in the eyes of fructivores as it is of the effects of interactions between sugar and other compounds in the chemistry of the fruit. Fruits that are not color-coded compete poorly on the shelves of nature's supermarket, but false advertising will be punished; the fruits that are ripe (full of nutrition) *and that advertise that fact* will sell better, but the advertising has to be tailored to the visual capabilities and proclivities of the target consumers.

In the beginning, colors were made to be seen by those who were made to see them. But this evolved gradually, by happenstance, taking serendipitous advantage of whatever materials lay at hand, occasionally exploding in a profusion of elaborations of a new Trick, and always tolerating a large measure of pointless variation *and* pointless (merely coincidental) constancy. These coincidental constancies often concerned "more fundamental" features of the physical world. Once there were creatures who could distinguish red from green berries, they could also distinguish red rubies from green emeralds, but this was just a coincidental bonus. The fact that there is a difference *in* color between rubies and emeralds can thus be considered to be a *derived* color phenomenon. Why is the sky blue? Because apples are red and grapes are purple, not the other way around.

It is a mistake to think that first there were colors — colored rocks, colored water, colored sky, reddish-orange rust and bright blue cobalt — and then Mother Nature came along and took advantage of those properties by using them to color-code things. It is rather that first there were various reflective properties of surfaces, reactive properties of photopigments, and so forth, and Mother Nature developed out of these raw materials efficient, mutually adjusted "color"-coding/"color"-vision systems, and among the properties that settled out of that design process are the properties we normal human beings call colors. If the blue of cobalt and the blue of a butterfly's wing happened to match (in normal human beings' vision) this is just a coincidence, a negligible side effect of the processes that brought color vision into existence and

thereby (as Locke himself might have acknowledged) baptized a certain curiously gerrymandered set of complexes of primary properties with the shared secondary property of producing a common effect in a set of normal observers.

"But still," you will want to object, "back before there were any animals with color vision, there were glorious red sunsets, and bright green emeralds!" Well, yes, you can say so, but then those very same sunsets were also garish, multicolored, and disgusting, rendered in colors we cannot see, and hence have no names for. That is, you will have to admit this, if there are or could *be* creatures on some planet whose sensory apparatus would be so affected by them. And for all we know, there are species somewhere who naturally see that there are two (or seventeen) different colors among a batch of emeralds *we* found to be indistinguishably green.

Many human beings are red-green colorblind. Suppose we all were; it would then be common knowledge that both rubies and emeralds were "gred" — after all, they look to normal observers just like other gred things: fire engines, well-watered lawns, apples ripe and unripe (Dennett, 1969). Were folks like us to come along, insisting that rubies and emeralds were in fact different colors, there would be no way to declare one of these color-vision systems "truer" than the other.

The philosopher Jonathan Bennett (1965) draws our attention to a case that makes the same point, more persuasively, in another sensory modality. The substance phenol-thio-urea, he tells us, tastes bitter to one-quarter of the human population and is utterly tasteless to the rest. Which way it tastes to you is genetically determined. Is phenol-thio-urea bitter or tasteless? By "eugenics" (controlled breeding) or genetic engineering, we might succeed in eliminating the genotype for finding phenol bitter. If we succeeded, phenol-thio-urea would then be *paradigmatically* tasteless, as tasteless as distilled water: tasteless to all normal human beings. If we performed the opposite genetic experiment, we could in time render phenol-thio-urea paradigmatically bitter. Now, before there were any human beings, was phenol-thio-urea *both* bitter and tasteless? It was chemically the same as it is now.

Facts about secondary qualities are inescapably linked to a reference class of observers, but there are weak and strong ways of treating the link. We may say that secondary qualities are *lovely* rather than suspect. Someone could be lovely who had never yet, as it happened, been observed by any observer of the sort who would find her lovely, but she could not — as a matter of logic — be a suspect until someone

actually suspected her of something. Particular instances of lovely qualities (such as the quality of loveliness) can be said to exist as Lockean dispositions prior to the moment (if any) where they exercise their power over an observer, producing the defining effect therein. Thus some unseen woman (self-raised on a desert island, I guess) could be genuinely lovely, having the dispositional power to affect normal observers of a certain class in a certain way, in spite of never having the opportunity to do so. But lovely qualities cannot be defined independently of the proclivities, susceptibilities, or dispositions of a class of observers, so it really makes no sense to speak of the existence of lovely properties in complete independence of the existence of the relevant observers. Actually, that's a bit too strong. Lovely qualities *would* not be defined — there would be no point in defining *them*, in contrast to all the other logically possible gerrymandered properties — independently of such a class of observers. So while it might be logically possible ("in retrospect," one might say) to gather color-property instances together by something like brute force enumeration, the reasons for singling out such properties (for instance, in order to explain certain causal regularities in a set of curiously complicated objects) depend on the existence of the class of observers.

Are sea elephants lovely? Not to us. It is hard to imagine an uglier creature. What makes a sea elephant lovely to another sea elephant is not what makes a woman lovely to a man, and to call some as-yet-unobserved woman lovely who, as it happens, would mightily appeal to sea elephants would be to abuse both her and the term. It is only by reference to human tastes, which are contingent and indeed idiosyncratic features of the world, that the property of loveliness (to-a-human-being) can be identified.

On the other hand, suspect qualities (such as the property of being a suspect) are understood in such a way as to presuppose that any instance of the property has already had its defining effect on at least one observer. You may be eminently worthy of suspicion — you may even be obviously guilty — but you can't be a suspect until someone actually suspects you. I am not claiming that colors are suspect qualities. Our intuition that the as-yet-unobserved emerald in the middle of the clump of ore is *already* green does not have to be denied. But I am claiming that colors are lovely qualities, whose existence, tied as it is to a reference class of observers, makes no sense in a world in which the observers have no place. This is easier to accept for some secondary qualities than for others. That the sulphurous fumes spewed

forth by primordial volcanos were yellow seems somehow more objective than that they stank, but so long as what we mean by "yellow" is what *we* mean by "yellow," the claims are parallel. For suppose some primordial earthquake cast up a cliff face exposing the stripes of hundreds of chemically different layers to the atmosphere. Were those stripes *visible*? We must ask to whom. Perhaps some of them would be visible to us and others not. Perhaps some of the invisible stripes would be visible to tetrachromat pigeons, or to creatures who saw in the infrared or ultraviolet part of the electromagnetic spectrum. For the same reason one cannot meaningfully ask whether the difference between emeralds and rubies is a visible difference without specifying the vision system in question.

Evolution softens the blow of the "subjectivism" or "relativism" implied by the fact that secondary qualities are lovely qualities. It shows that the absence of "simple" or "fundamental" commonalities in things that are all the same color is not an earmark of total illusion, but rather, a sign of a widespread tolerance for "false positive" detections of the ecological properties that really matter.[2] The basic categories of our color spaces (and of course our odor spaces and sound spaces, and all the rest) are shaped by selection pressures, so that in general it makes sense to ask what a particular discrimination or preference is for. There are reasons why we shun the odors of certain things and seek out others, why we prefer certain colors to others, why some sounds bother us more, or soothe us more. They may not always be our reasons, but rather the reasons of distant ancestors, leaving their fossil traces in the built-in biases that innately shape our quality spaces. But as good Darwinians, we should also recognize the possibility — indeed, the necessity — of other, nonfunctional biases, distributed haphazardly

2. Philosophers are currently fond of the concept of *natural kinds*, reintroduced to philosophy by Quine (1969), who may now regret the way it has become a stand-in for the dubious but covertly popular concept of *essences*. "Green things, or at least green emeralds, are a kind," Quine observes (p. 116), manifesting his own appreciation of the fact that while emeralds may be a natural kind, green things are probably not. The present discussion is meant to forestall one of the tempting mistakes of armchair naturalism: the assumption that whatever nature makes is a natural kind. Colors are not "natural kinds" precisely *because* they are the product of biological evolution, which has a tolerance for sloppy boundaries when making categories that would horrify any philosopher bent on good clean definitions. If some creature's life depended on lumping together the moon, blue cheese, and bicycles, you can be pretty sure that Mother Nature would find a way for it to "see" these as "intuitively just the same kind of thing."

through the population in genetic variation. In order for selection pressure to differentially favor those who exhibit a bias against *F* once *F* becomes ecologically important, there has to have been pointless (not-yet-functional) variation in "attitude toward *F*" on which selection can act. For example, if eating tripe were to spell prereproductive doom in the future, only those of us who were "naturally" (and heretofore pointlessly) disposed against eating tripe would have an advantage (perhaps slight to begin with, but soon to be explosive, if conditions favored it). So it doesn't follow that if you find something (e.g., broccoli) indescribably and ineffably awful, there is a reason for this. Nor does it follow that you are defective if you disagree with your peers about this. It may just be one of the innate bulges in your quality space that has, as of yet, no functional significance at all. (And for your sake, you had better hope that if it ever does have significance, it is because broccoli has suddenly turned out to be bad for us.)

These evolutionary considerations go a long way to explaining why secondary qualities turn out to be so "ineffable," so resistant to definition. Like the shape property *M* of the Rosenbergs' piece of Jell-O box, secondary qualities are extremely resistant to straightforward definition. It is of the essence of the Rosenbergs' trick that we cannot replace our dummy predicate *M* with a longer, more complex, but accurate and exhaustive description of the property, for if we could, we (or someone else) could use that description as a recipe for producing another instance of *M* or another *M*-detector. Our secondary quality detectors were not specifically designed to detect only hard-to-define properties, but the result is much the same. As Akins (1989) observes, it is not *the point* of our sensory systems that they should detect "basic" or "natural" properties of the environment, but just that they should serve our "narcissistic" purposes in staying alive; nature doesn't build epistemic engines.

The only *readily available* way of saying just what shape property *M* is is just to point to the *M*-detector and say that *M* is the shape property detected by this thing here. The same predicament naturally faces anyone trying to say what property someone detects (or misdetects) when something "looks the way it looks to him." So now we can answer the question with which this section began: What property does Otto judge something to have when he judges it to be pink? The property he calls pink. And what property is that? It's hard to say, but this should not embarrass us, because we can say why it's hard to say. The best we can do, practically, when asked what surface properties we detect with color vision, is to say, uninformatively, that we detect the prop-

erties we detect. If someone wants a more informative story about those properties, there is a large and rather incompressible literature in biology, neuroscience, and psychophysics to consult. And Otto can't say anything more about the property he calls pink by saying "It's *this*!" (taking himself to be pointing "inside" at a private, phenomenal property of his experience). All that move accomplishes (at best) is to point to his own idiosyncratic color-discrimination state, a move that is parallel to holding up a piece of Jell-O box and saying that it detects this shape property. Otto points to his discrimination-device, perhaps, but not to any quale that is exuded by it, or worn by it, or rendered by it, when it does its work. There are no such things.

> But still [Otto insists], you haven't yet said why pink should look like *this!*

Like what?

> Like *this*. Like the particularly ineffable, wonderful, intrinsic pinkness that I am right now enjoying. *That* is not some indescribably convoluted surface reflectance property of external objects.

I see, Otto, that you use the term *enjoying*. You are not alone. Often, when an author wants to stress that the topic has turned from (mere) neuroanatomy to experience, (mere) psychophysics to consciousness, (mere) information to qualia, the word "enjoy" is ushered onto the stage.

3. ENJOYING OUR EXPERIENCES

> But Dan, qualia are what make life worth living!
>
> WILFRID SELLARS (over a fine bottle of Chambertin, Cincinnati, 1971)

> If what I want when I drink fine wine is information about its chemical properties, why don't I just read the label?
>
> SYDNEY SHOEMAKER, Tufts Colloquium, 1988

Some colors were made for liking, and so were some smells and tastes. And other colors, smells, and tastes, were made for disliking. To put the same point more carefully, it is no accident that we (and other creatures who can detect them) like and dislike colors, smells, tastes, and other secondary qualities. Just as we are the inheritors of

evolved vertical symmetry detectors in our visual systems for alerting us (like our ancestors) to the ecologically significant fact that another creature is looking at us, so we are the inheritors of evolved quality-detectors that are not disinterested reporters, but rather warners and beckoners, sirens in both the fire-engine sense and the Homeric sense.

As we saw in chapter 7, on evolution, these native alarmists have subsequently been coopted in a host of more complicated organizations, built from millions of associations, and shaped, in the human case, by thousands of memes. In this way the brute come-and-get-it appeal of sex and food, and the brute run-for-your-life aversion of pain and fear get stirred together in all sorts of piquant combinations. When an organism discovers that it pays to attend to some feature of the world *in spite of* its built-in aversion to doing that, it must construct some countervailing coalition to keep aversion from winning. The resulting semistable tension can then itself become an acquired taste, to be sought out under certain conditions. When an organism discovers that it must smother the effects of certain insistent beckoners if it is to steer the proper course, it may cultivate a taste for whatever sequences of activity it can find that tend to produce the desired peace and quiet. In such a way could we come to love spicy food that burns our mouths (Rozin, 1982), deliciously "discordant" music, and both the calm, cool realism of Andrew Wyeth and the unsettling, hot expressionism of Willem de Kooning. Marshall McLuhan (1967) proclaimed that the medium is the message, a half-truth that is truer perhaps in the nervous system than in any other forum of communication. What we want when we sip a great wine is not, indeed, the information about its chemical contents; what we want is *to be informed* about its chemical contents in our favorite way. And our preference is *ultimately* based on the biases that are still wired into our nervous systems though their ecological significance may have lapsed eons ago.

This fact has been largely concealed from us by our own technology. As the psychologist Nicholas Humphrey notes,

> As I look around the room I'm working in, man-made colour shouts back at me from every surface: books, cushions, a rug on the floor, a coffee-cup, a box of staples — bright blues, reds, yellows, greens. There is as much colour here as in any tropical forest. Yet while almost every colour in the forest would be meaningful, here in my study almost nothing is. Colour anarchy has taken over. [1983, p. 149]

Consider, for instance, the curious fact that monkeys don't like red light. Given a choice, rhesus monkeys show a strong preference for the blue-green end of the spectrum, and get agitated when they have to endure periods in red environments (Humphrey, 1972, 1973, 1983; Humphrey and Keeble, 1978). Why should this be? Humphrey points out that red is always used to alert, the ultimate color-coding color, but for that very reason ambiguous: the red fruit may be good to eat, but the red snake or insect is probably advertising that it is poisonous. So "red" sends mixed messages. But why does it send an "alert" message in the first place? Perhaps because it is the strongest available contrast with the ambient background of vegetative green or sea blue, or — in the case of monkeys — because red light (red to reddish-orange to orange light) is the light of dusk and dawn, the times of day when virtually all the predators of monkeys do their hunting.

The affective or emotional properties of red are not restricted to rhesus monkeys. All primates share these reactions, including human beings. If your factory workers are lounging too long in the rest rooms, painting the walls of the rest rooms red will solve that problem — but create others (see Humphrey, forthcoming). Such "visceral" responses are not restricted to colors, of course. Most primates raised in captivity who have never seen a snake will make it unmistakably clear that they loathe snakes the moment they see one, and it is probable that the traditional human dislike of snakes has a biological source that explains the biblical source, rather than the other way around.[3] That is, our genetic heritage queers the pitch in favor of memes for snake-hating.

Now here are two different explanations for the uneasiness most of us feel (even if we "conquer" it) when we see a snake:

(1) Snakes evoke in us a particular intrinsic snake-yuckiness quale when we look at them, and our uneasiness is a reaction to that quale.
(2) We find ourselves less than eager to see snakes because of innate biases built into our nervous systems. These favor the release of adrenaline, bring fight-or-flight routines on line, and,

3. The primatologist Sue Savage-Rumbaugh has informed me that laboratory-raised bonobos, or pygmy chimps, show no signs of an innate dislike of snakes, unlike chimpanzees.

> by activating various associative links, call a host of scenarios into play involving danger, violence, damage. The original primate aversion is, in us, transformed, revised, deflected in a hundred ways by the memes that have exploited it, coopted it, shaped it. (There are many different levels at which we could couch an explanation of this "functionalist" type. For instance, we could permit ourselves to speak more casually about the power of snake-perceptions to produce anxieties, fears, anticipations of pain, and the like, but that might be seen as "cheating" so I am avoiding it.)

The trouble with the first sort of explanation is that it only seems to be an explanation. The idea that an "intrinsic" property (of occurrent pink, of snake-yuckiness, of pain, of the aroma of coffee) could *explain* a subject's reactions to a circumstance is hopeless — a straightforward case of a *virtus dormitiva* (see page 63). Convicting a theory of harboring a vacuous *virtus dormitiva* is not that simple, however. Sometimes it makes perfectly good sense to posit a temporary *virtus dormitiva*, pending further investigation. Conception is, by definition we might say, the cause of pregnancy. If we had no other way of identifying conception, telling someone she got pregnant because she conceived would be an empty gesture, not an explanation. But once we've figured out the requisite mechanical theory of conception, we can see *how* conception is the cause of pregnancy, and informativeness is restored. In the same spirit, we might identify qualia, by definition, as the proximal causes of our enjoyment and suffering (roughly put), and then proceed to discharge our obligations to inform by pursuing the second style of explanation. But curiously enough, qualophiles (as I call those who still believe in qualia) will have none of it; they insist, like Otto, that qualia "reduced" to mere complexes of mechanically accomplished dispositions to react are not the qualia they are talking about. *Their* qualia are something different.

> Consider [says Otto] the way the pink ring seems to me right now, at this very moment, in isolation from all my dispositions, past associations and future activities. *That*, the purified, isolated *way it is with* me in regards to color at this moment — that is my pink quale.

Otto has just made a mistake. In fact, this is the big mistake, the source of all the paradoxes about qualia, as we shall see. But before exposing the follies of taking this path, I want to demonstrate some of the positive

benefits of the path that Otto shuns: the "reductionist" path of identifying "the way it is with me" with the sum total of all the idiosyncratic reactive dispositions inherent in my nervous system as a result of my being confronted by a certain pattern of stimulation.

Consider what it must have been like to be a Leipzig Lutheran churchgoer in, say, 1725, hearing one of J. S. Bach's chorale cantatas in its premier performance. (This exercise in imagining *what it is like* is a warm-up for chapter 14, where we will be concerned with consciousness in other animals.) There are probably no significant biological differences between us today and German Lutherans of the eighteenth century; we are the same species, and hardly any time has passed. But, because of the tremendous influence of culture — the memosphere — our psychological world is quite different from theirs, in ways that would have a noticeable impact on our respective experiences when hearing a Bach cantata for the first time. Our musical imagination has been enriched and complicated in many ways (by Mozart, by Charlie Parker, by the Beatles), but also it has lost some powerful associations that Bach could count on. His chorale cantatas were built around chorales, traditional hymn melodies that were deeply familiar to his churchgoers and hence provoked waves of emotional and thematic association as soon as their traces or echoes appeared in the music. Most of us today know these chorales only from Bach's settings of them, so when we hear them, we hear them with different ears. If we want to imagine what it was like to be a Leipzig Bach-hearer, it is not enough for us to hear the same tones on the same instruments in the same order; we must also prepare ourselves somehow to respond to those tones with the same heartaches, thrills, and waves of nostalgia.

It is not utterly impossible to prepare ourselves in these ways. A music scholar who carefully avoided all contact with post-1725 music and familiarized himself intensively with the traditional music of that period would be a good first approximation. More important, as these observations show, it is not impossible to know in just what ways we would have to prepare ourselves whether or not we cared to go to all the trouble. So we could know what it was like "in the abstract" so to speak, and in fact I've just told you: the Leipzigers, hearing the chorale cantatas, were reminded of all the associations that already flavored their recognition of the chorale melodies. It is easy enough to imagine what *that* must have been like for them — though with variations drawn from our own experience. We can imagine what it would be like to hear Bach's setting of familiar Christmas carols, for instance, or "Home on the Range." We can't do the job precisely, but only because we can't

forget or abandon all that we know that the Leipzigers didn't know.

To see how crucial this excess baggage of ours is, imagine that musicologists unearthed a heretofore unknown Bach cantata, definitely by the great man, but hidden in a desk and probably never yet heard even by the composer himself. Everyone would be aching to hear it, to experience for the first time the "qualia" that the Leipzigers would have known, had they only heard it, but this turns out to be impossible, for the main theme of the cantata, by an ugly coincidence, is the first seven notes of "Rudolph the Red-Nosed Reindeer"! We who are burdened with that tune would never be able to hear Bach's version as he intended it or as the Leipzigers would have received it.

A clearer case of imagination-blockade would be hard to find, but note that it has nothing to do with biological differences or even with "intrinsic" or "ineffable" properties of Bach's music. The reason we couldn't imaginatively relive in detail (and accurately) the musical experience of the Leipzigers is simply that we would have to take ourselves along for the imaginary trip, and we know too much. But if we want, we can carefully list the differences between our dispositions and knowledge and theirs, and by comparing the lists, come to appreciate, in *whatever detail we want*, the differences between what it was like to be them listening to Bach, and what it is like to be us. While we might lament that inaccessibility, at least we could understand it. There would be no mystery left over; just an experience that could be described quite accurately, but not directly enjoyed unless we went to ridiculous lengths to rebuild our personal dispositional structures.

Qualophiles, however, have resisted this conclusion. It has seemed to them that even though such an investigation as we have just imagined might settle *almost* all the questions we had about what it was like to be the Leipzigers, there would have to be an ineffable residue, *something* about what it was like for the Leipzigers that no further advances in merely "dispositional" and "mechanistic" knowledge could reduce to zero. That is why qualia have to be invoked by qualophiles as *additional* features, over and above and strictly independent of the wiring that determines withdrawal, frowning, screaming, and other "mere behaviors" of disgust, loathing, fear. We can see this clearly if we revert to our example of colors.

Suppose we suggest to Otto that what made his "occurrent pink" the particular tantalizing experience that he enjoyed was simply the sum total of all the innate and learned associations and reactive dispositions triggered by the particular way he was (mis)informed by his eyes:

What qualia are, Otto, are just those complexes of dispositions. When you say "This is my quale," what you are singling out, or referring to, whether you realize it or not, is your idiosyncratic complex of dispositions. You seem to be referring to a private, ineffable something-or-other in your mind's eye, a private shade of homogeneous pink, but this is just how it seems to you, not how it is. That "quale" of yours is a character in good standing in the fictional world of your heterophenomenology, but what it turns out to be in the *real* world in your brain is just a complex of dispositions.

> That cannot be all there is to it [Otto replies, taking the fatal step in the qualophile tradition], for while that complex of mere dispositions might be the basis or source, somehow, for my particular quale of pink, they could all be changed without changing my intrinsic quale, or my intrinsic quale could change, without changing that manifold of mere dispositions. For instance, my qualia could be *inverted* without inverting all my dispositions. I could have all the reactivities and associations that I now have for green to the accompaniment of the *quale* I now have for red, and vice versa.

4. A PHILOSOPHICAL FANTASY: INVERTED QUALIA

The idea of the possibility of such "inverted qualia" is one of philosophy's most virulent memes. Locke discussed it in his *Essay Concerning Human Understanding* (1690), and many of my students tell me that as young children they hit upon the same idea for themselves, and were fascinated by it. The idea seems to be transparently clear and safe:

> There are the ways things look to me, and sound to me, and smell to me, and so forth. That much is obvious. I wonder, though, if the ways things appear to me are the same as the ways things appear to other people.

Philosophers have composed many different variations on this theme, but the classic version is the interpersonal version: How do I know that you and I see the same subjective color when we look at something? Since we both learned our color words by being shown public colored objects, our verbal behavior will match even if we experience entirely different subjective colors — even if the way red things look to me is the way green things look to you, for instance. We would call the same

public things "red" and "green" even if our private experiences were "the opposite" (or just different).

Is there any way to tell whether this is the case? Consider the hypothesis that red things look the same to you and me. Is this hypothesis both irrefutable and unconfirmable? Many have thought so, and some have concluded that for just that reason it is one sort of nonsense or another, in spite of its initial appeal to common sense. Others have wondered if technology might come to the rescue and confirm (or disconfirm) the interpersonal inverted spectrum hypothesis. The science-fiction movie *Brainstorm* (not, I hasten to say, a version of my book *Brainstorms*) featured just the right imaginary device: Some neuroscientific apparatus fits on your head and feeds your visual experience into my brain via a cable. With eyes closed I accurately report everything you are looking at, except that I marvel at how the sky is yellow, the grass red, and so forth. If we had such a machine, couldn't such an experiment with it confirm, empirically, the hypothesis that our qualia were different? But suppose the technician pulls the plug on the connecting cable, inverts it 180 degrees, reinserts it in the socket, and I now report the sky is blue, the grass green, and so forth. Which would be the "right" orientation of the plug? Designing and building such a device — supposing for the moment that it would be possible — would require that its "fidelity" be tuned or calibrated by the normalization of the two subjects' reports, so we would be right back at our evidential starting point. Now one might try to avert this conclusion with further elaborations, but the consensus among the qualophiles is that this is a lost cause; there seems to be general agreement that the moral of *this* thought experiment is that no intersubjective comparison of qualia would be possible, even with perfect technology. This does provide support, however, for the shockingly "verificationist" or "positivistic" view that the very idea of inverted qualia is nonsense — and hence that the very idea of qualia is nonsense. As the philosopher Ludwig Wittgenstein put it, using his famous "beetle in the box" analogy,

> The thing in the box has no place in the language-game at all; not even as a something; for the box might even be empty. — No, one can "divide through" by the thing in the box; it cancels out, whatever it is. [1953, p. 100]

But just what does this mean? Does it mean that qualia are real but ineffective? Or that there aren't any qualia after all? It still seemed obvious to most philosophers who thought about it that qualia were

real, even if a difference in qualia would be a difference that couldn't be detected in any way. That's how matters stood, uneasily, until someone dreamt up the presumably improved version of the thought experiment: the intrapersonal inverted spectrum. The idea seems to have occurred to several people independently (Gert, 1965; Putnam, 1965; Taylor, 1966; Shoemaker, 1969; Lycan, 1973). In this version, the experiences to be compared are all in one mind, so we don't need the hopeless Brainstorm machine.

> You wake up one morning to find that the grass has turned red, the sky yellow, and so forth. No one else notices any color anomalies in the world, so the problem must be in you. You are entitled, it seems, to conclude that you have undergone visual color qualia inversion. How did it happen? It turns out that while you slept, evil neurosurgeons switched all the wires — the neurons — leading from the color-sensitive cone cells in your retinas.

So far, so good. The effect on you would be startling, maybe even terrifying. You would certainly be able to detect that the way things looked to you now was very different, and we would even have a proper scientific explanation of why this was: The neuron clusters in the visual cortex that "care about" color, for instance, would be getting their stimulation from a systematically shifted set of retinal receptors. So half the battle is won, it seems: A difference in qualia would be detectable after all, if it were a difference that developed rather swiftly in a single person.[4] But this is only half the battle, for the imagined neurosurgical prank has also switched all your reactive dispositions; not only do you say your color experiences have all been discombobulated, but your nonverbal color-related behavior has been inverted as well. The edginess you used to exhibit in red light you now exhibit in green light, and you've lost the fluency with which you used to rely on various color-coding schemes in your life. (If you play basketball for the Boston Celtics, you keep passing the ball mistakenly to the guys in the red uniforms.)

What the qualophile needs is a thought experiment that demon-

4. The suddenness would be important, since if it happened very gradually, you might not be able to notice. As Hardin (1990) has pointed out, the gradual yellowing of your lenses with age slowly shifts your sense of the primary colors; shown a color wheel and asked to point at pure red (red with no orange or purple in it), where on the continuum you point is partly a function of age.

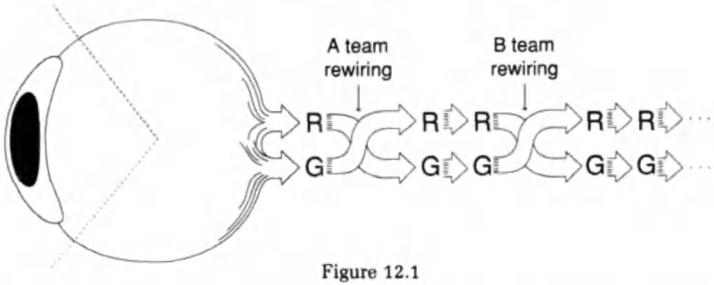

Figure 12.1

strates that the-way-things-look can be independent of all these reactive dispositions. So we have to complicate the story with a further development; we must describe something happening that undoes the switch in reactive dispositions while leaving the switched "qualia" intact. Here is where the literature lurches into ever more convoluted fantasies, for no one thinks for a moment that the-way-things-look is ever *actually* divorced from the subject's reactive dispositions; it is just that this is deemed an important *possibility in principle* by the qualophiles. To show this, they need to describe a possible case, however outlandish, in which it would be obvious that this detachment was actual. Consider a story that *won't* work:

> One night while you sleep, evil neurosurgeons switch all the wires from the cone cells (just as before), and then, later the same busy night, another team of neurosurgeons, the B team, comes along and performs a *complementary* rewiring a little farther up on the optic nerve.

This restores all the old reactive dispositions (we can presume), but, alas, it also restores the old qualia. The cells in the cortex that "care about" color, for instance, will now be getting their original signals again, thanks to the speedy undoing of the damage by the B team. The second switcheroo happened too early, it seems; it happened *on the way up* to conscious experience. So we'll have to tell the story differently, with the second switcheroo happening later, *after* the inverted qualia have taken their bow in consciousness, but *before* any of the inverted reactions to them can set in. But is this possible? Not if the arguments for the Multiple Drafts model are correct. There is no line that can be drawn across the causal "chain" from eyeball through consciousness to subsequent behavior such that all reactions to x happen after it and consciousness of x happens before it. This is because it is not a simple causal chain, but a causal network, with multiple paths on which Multiple Drafts are being edited simultaneously and

semi-independently. The qualophile's story would make sense if there were a Cartesian Theater, a special place in the brain where the conscious experience happened. If there were such a place, we could bracket it with the two switcheroos, leaving inverted qualia in the Theater, while keeping all the reactive dispositions normalized. Since there is no such Cartesian Theater, however, the thought experiment doesn't make sense. There is no coherent way to tell the necessary story. There is no way to isolate the properties *presented* in consciousness from the brain's multiple reactions to its discriminations, because there is no such additional presentation process.

In the literature on the inverted spectrum, the second switcheroo is often supposed to be accomplished not by surgery but by gradual adaptation by the subject to the new regime of experiences. This makes superficial sense; people can adapt amazingly well to bizarre displacements of their senses. There have been many visual field inversion experiments in which subjects wear goggles that turn everything upside down — by turning the retinal image right side up! (E.g., Stratton, 1896; Kohler, 1961; Welch, 1978, provides a good summary; see also Cole, 1990.) After several days of constantly wearing inverting goggles of one sort or another (it makes a difference — some varieties had a wide field of view, and others gave the viewers a sort of tunnel vision), subjects often make an astonishingly successful adaptation. In Ivo Kohler's film of his experiments in Innsbruck, we see two of his subjects, comically helpless when they first put on the goggles, skiing downhill and riding bicycles through city traffic, still wearing the inverting goggles and apparently completely adapted to them.

So let's suppose that you gradually adapt to the surgical inversion of your color vision. (Why you would want to adapt, or would have to adapt, is another matter, but we may as well concede the point to the qualophiles, to hasten their demise.) Now some adaptations would at first be clearly *post-experiential*. We may suppose that the clear sky would still *look* yellow to you, but you would start calling it blue to get in step with your neighbors. Looking at a novel object might cause momentary confusion: "It's gr— I mean red!" What about your edginess in green light — would it still show up as an abnormality in your galvanic skin response? For the sake of the argument, the qualophile has to imagine, however unlikely this might be, that *all* your reactive dispositions adapt, leaving behind only the residue of the still-inverted qualia, so for the sake of argument, let's concede that the most fundamental and innate biases in your quality spaces also "adapt" — this is preposterous, but there is worse to come.

In order to tell the necessary story, the qualophile must suppose that eventually all these adaptations become second nature — swift and unstudied. (If they didn't become second nature, there would be leftover reactive dispositions that would be still different, and the argument requires that these all be ironed out.) So be it. Now, assuming that *all* your reactive dispositions are restored, what is your intuition about your qualia? Are they still inverted or not?

It is legitimate to pass at this point, on the grounds that after being asked to tolerate so many dubious assumptions for the sake of argument, you either come up empty — no intuition bubbles up at all — or you find yourself mistrustful of whatever intuition does strike you. But perhaps it does seem quite obvious to you that your qualia would still be inverted. But why? What in the story has led you to see it this way? Perhaps, even though you have been following directions, you have innocently added some further assumptions not demanded by the story, or failed to notice certain possibilities not ruled out by the story. I suggest that the most likely explanation for your intuition that, in this imagined instance, you would still have "inverted qualia" is that you are making the additional, and unwarranted, assumption that all the adaptation is happening on the "post-experiential side."

It could be, though, couldn't it, that the adaptation was accomplished on the upward path? When you first put on heavily tinted goggles, you won't see any color at all — or at least the colors you see are weird and hard-to-distinguish colors — but after wearing them for a while, surprising normal color vision returns. (Cole, 1990, draws philosophers' attention to these effects, which you can test for yourself with army-surplus infrared sniper goggles.) Perhaps, not knowing this surprising fact, it just never occurred to you that you *might* adapt to the surgery in much the same way. We could have highlighted this possibility in the thought experiment, by adding a few details:

> . . . And as the adaptation proceeded, you often found to your surprise that the colors of things didn't seem so strange after all, and sometimes you got confused and made double corrections. When asked the color of a novel object you said "It's gr— , no red — no, it *is* green!"

Told this way, the story might make it seem "obvious" that the color qualia themselves had adapted, or been reinverted. But in any case, you may now think, it has to be one way or the other. There couldn't be a case where it wasn't perfectly obvious which sort of adjustment you had made! The unexamined assumption that grounds *this* convic-

tion is that all adaptations can be categorized as either pre-experiential or post-experiential (Stalinesque or Orwellian). At first this may seem to be an innocent assumption, since extreme cases are easy to classify. When the brain compensates for head and eye motions, producing a stable visual world "in experience," this is surely a pre-experiential cancellation of effects, an adaptation on the upward path to consciousness. And when you imagine making peripheral ("late") compensations in color-word choosing ("It's gr— I mean *red!*") this is obviously a post-experiential, merely behavioral adjustment. It stands to reason then, doesn't it, that when *all* the adaptations have been made, either they leave the subjective color (the color "in consciousness") inverted or they don't? Here's how we would tell: Add up the switcheroos on the upward path; if there are an even number — as in the Team B handiwork — the qualia are normalized, and if odd, the qualia are still inverted. Nonsense. Recall the Neo-Laffer curve in chapter 5. It is not at all a logical or geometric necessity that there be a single value of a discriminated variable that can be singled out as *the* value of the variable "in consciousness."

We can demonstrate this with a little fantasy of our own, playing by the qualophile's rules. Suppose that presurgically a certain shade of blue tended to remind you of a car in which you once crashed, and hence was a color to be shunned. At first, postsurgically, you have no negative reactions to things of that color, finding them an innocuous and unmemorable yellow, let's suppose. After your complete adaptation, however, you again shun things of that shade of blue, *and it is because they remind you of that crash.* (If they didn't, this would be an unadapted reactive disposition.) But if we ask you whether this is because, as you remember the crash, the car was yellow — just like the noxious object before you now — or because, as you remember the crash, the car was blue — just like the noxious object before you now, you really shouldn't be able to answer. Your verbal behavior will be totally "adapted"; your immediate, second-nature answer to the question: "What color was the car you crashed?" is "blue" and you will unhesitatingly call the noxious object before you blue as well. Does that entail that you have *forgotten* the long training period?

No. We don't need anything so dramatic as amnesia to explain your inability to answer, for we have plenty of everyday cases in which the same phenomenon arises. Do you like beer? Many people who like beer will acknowledge that beer is an acquired taste. One gradually trains oneself — or just comes — to enjoy that flavor. What flavor? The flavor of the first sip?

> No one could like *that* flavor [an experienced beer drinker might retort]. Beer tastes different to the experienced beer drinker. If beer went on tasting to me the way the first sip tasted, I would never have gone on drinking beer! Or, to put the same point the other way around, if my first sip of beer had tasted to me the way my most recent sip just tasted, I would never have had to acquire the taste in the first place! I would have loved the first sip as much as the one I just enjoyed.

If this beer drinker is right, then beer is not an acquired taste. No one comes to enjoy *the way the first sip tasted*. Instead, *the way beer tastes* to them gradually changes. Other beer drinkers might insist that, no, beer *did* taste to them now the way it always did, only now they like *that very taste*. Is there a real difference? There is a difference in heterophenomenology, certainly, and the difference needs to be explained. It *could* be that the different convictions spring from genuine differences in discriminative capacity of the following sort: in the first sort of beer drinker the "training" has changed the "shape" of the quality space for tasting, while in the second sort the quality space remains roughly the same, but the "evaluation function" over that space has been revised. Or it *could* be that some or even all of the beer drinkers are kidding themselves (like those who insist that the high-resolution Marilyns are all really *there* in the background of their visual field). We have to look beyond the heterophenomenological worlds to the actual happenings in the head to see whether there is a truth-preserving (if "strained") interpretation of the beer drinkers' claims, and *if* there is, it will only be because we decide to reduce "the way it tastes" to one complex of reactive dispositions or another (Dennett, 1988a). We would have to "destroy" qualia in order to "save" them.

So if a beer drinker furrows his brow and gets a deadly serious expression on his face and says that what he is referring to is "*the way the beer tastes to me right now*," he is definitely kidding himself if he thinks he can *thereby* refer to a quale of his acquaintance, a subjective state that is independent of his changing reactive attitudes. It may seem to him that he can, but he can't.[5]

And by the same token, in the imagined case of being reminded

---

5. "The very fact that we should so much like to say: 'This is the important thing' — while we point privately to the sensation — is enough to shew how much we are inclined to say something which gives no information." Wittgenstein (1953), i298.

of the car crash by the blue object, you would be kidding yourself if you thought you could tell, from *the way the object looks to you*, whether it was "intrinsically" the same as the way the car looked to you when you crashed. This is enough to undercut the qualophile's thought experiment, for the goal was to describe a case in which it was *obvious* that the qualia would be inverted while the reactive dispositions would be normalized. The *assumption* that one could just tell is question-begging, and without the assumption, there is no argument, but just an intuition pump — a story that cajoles you into declaring your gut intuition without giving you a good reason for it.

Question-begging or not, it may *still* seem just plain obvious that "the subjective colors you would be seeing things to be" would *have* to be "one way or the other." This just shows the powerful gravitational force that the Cartesian Theater exerts on our imaginations. It may help to break down the residual attractiveness of this idea if we consider further the invited parallel with image-inverting goggles. When the adaptations of the subjects wearing these goggles have become so second nature that they can ride bicycles and ski, the natural (but misguided) question to ask is this: Have they adapted *by turning their experiential world back right side up*, or *by getting used to their experiential world being upside down*? And what do they say? They say different things, which correlate roughly with how complete their adaptation was. The more complete it was, the more the subjects dismiss the question as improper or unanswerable. This is just what the Multiple Drafts theory demands: Since there are a host of discriminations and reactions that need to be adjusted, scattered around in the brain, some of them dealing with low-level "reflexes" (such as ducking the right way when something looms at you) and others dealing with focally attended deliberate actions, it is not surprising that as the adaptations in this patchwork accumulate, subjects should lose all conviction of whether to say "things look the way they used to look" instead of "things still look different, but I'm getting used to it." In some ways things look the same to them (as judged by their reactions), in other ways things look different (as judged by other reactions). If there were a single representation of visuo-motor space through which all reactions to visual stimuli had to be channeled, *it* would have to be "one way or the other," perhaps, but there is no such single representation. *The* way things look to them is composed of many partly independent habits of reaction, not a single intrinsically right-side-up or upside-down picture in the head. All that matters is the fit between the input and the

output, and since this is accomplished in many different places with many different and largely independent means, there is just no saying what "counts" as "my visual field is still upside down."

The same is true of "qualia" inversion. The idea that it is something *in addition* to the inversion of all one's reactive dispositions, so that, if they were renormalized the inverted qualia would remain, is simply part of the tenacious myth of the Cartesian Theater. This myth is celebrated in the elaborate thought experiments about spectrum inversion, but to celebrate is not to support or prove. *If* there are no qualia over and above the sum total of dispositions to react, the idea of holding the qualia constant while adjusting the dispositions is self-contradictory.

5. "EPIPHENOMENAL" QUALIA?

There is another philosophical thought experiment about our experience of color that has proven irresistible: Frank Jackson's (1982) much-discussed case of Mary, the color scientist who has never seen colors. Like a good thought experiment, its point is immediately evident to even the uninitiated. In fact it is a bad thought experiment, an intuition pump that actually encourages us to misunderstand its premises!

> Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black-and-white room *via* a black-and-white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like *red*, *blue*, and so on. She discovers, for example, just which wavelength combinations from the sky stimulate the retina, and exactly how this produces *via* the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence "The sky is blue." . . . What will happen when Mary is released from her black-and-white room or is given a color television monitor? Will she *learn* anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. Ergo there is more to have than that, and Physicalism is false. . . . [p. 128]

The point could hardly be clearer. Mary has had no experience of color at all (there are no mirrors to look at her face in, she's obliged to wear black gloves, etc., etc.), and so, at that special moment when her captors finally let her come out into the colored world which she knows only by description (and black-and-white diagrams), "it seems just obvious," as Jackson says, that she will learn something. Indeed, we can all vividly imagine her, seeing a red rose for the first time and exclaiming, "So *that*'s what red looks like!" And it may also occur to us that if the first colored things she is shown are, say, unlabeled wooden blocks, and she is told only that one of them is red and the other blue, she won't have the faintest idea which is which until she somehow learns which color words go with her newfound experiences.

That is how almost everyone imagines this thought experiment — not just the uninitiated, but the shrewdest, most battle-hardened philosophers (Tye, 1986; Lewis, 1988; Loar, 1990; Lycan, 1990; Nemirov, 1990; Harman, 1990; Block, 1990; van Gulick, 1990). Only Paul Churchland (1985, 1990) has offered any serious resistance to the *image*, so vividly conjured up by the thought experiment, of Mary's dramatic discovery. The image is wrong; if that is the way you imagine the case, you are simply not following directions! The reason no one follows directions is because what they ask you to imagine is so preposterously immense, you can't even try. The crucial premise is that "She has *all* the physical information." That is not readily imaginable, so no one bothers. They just imagine that she knows lots and lots — perhaps they imagine that she knows everything that anyone knows *today* about the neurophysiology of color vision. But that's just a drop in the bucket, and it's not surprising that Mary would learn something if *that* were all she knew.

To bring out the illusion of imagination here, let me continue the story in a surprising — but legitimate — way:

> And so, one day, Mary's captors decided it was time for her to see colors. As a trick, they prepared a bright blue banana to present as her first color experience ever. Mary took one look at it and said "Hey! You tried to trick me! Bananas are yellow, but this one is blue!" Her captors were dumfounded. How did she do it? "Simple," she replied. "You have to remember that I know *everything* — absolutely everything — that could ever be known about the physical causes and effects of color vision. So of course before you brought the banana in, I had already written down, in exquisite detail, exactly what physical impression a yellow object

> or a blue object (or a green object, etc.) would make on my nervous system. So I already knew exactly what *thoughts* I would have (because, after all, the "mere disposition" to think about this or that is not one of your famous qualia, is it?). I was not in the slightest surprised by my experience of blue (what surprised me was that you would try such a second-rate trick on me). I realize it is *hard for you to imagine* that I could know so much about my reactive dispositions that the way blue affected me came as no surprise. Of course it's hard for you to imagine. It's hard for anyone to imagine the consequences of someone knowing absolutely everything physical about anything!"

Surely I've cheated, you think. I must be hiding some impossibility behind the veil of Mary's remarks. Can you prove it? My point is not that my way of telling the rest of the story proves that Mary doesn't learn anything, but that the usual way of imagining the story doesn't prove that she does. It doesn't prove anything; it simply pumps the intuition that she does ("it seems just obvious") by lulling you into imagining something other than what the premises require.

It is of course true that in any realistic, readily imaginable version of the story, Mary would come to learn something, but in any realistic, readily imaginable version she might know a lot, but she would not know everything physical. Simply imagining that Mary knows a lot, and leaving it at that, is not a good way to figure out the implications of her having "all the physical information" — any more than imagining she is filthy rich would be a good way to figure out the implications of the hypothesis that she owned everything. It may help us imagine the extent of the powers her knowledge gives her if we begin by enumerating a few of the things she obviously knows in advance. She knows black and white and shades of gray, and she knows the difference between the color of any object and such surface properties as glossiness versus matte, and she knows all about the difference between luminance boundaries and color boundaries (luminance boundaries are those that show up on black-and-white television, to put it roughly). And she knows precisely which effects — described in neurophysiological terms — each particular color will have on her nervous system. So the only task that remains is for her to figure out a way of identifying those neurophysiological effects "from the inside." You may find you can readily imagine her making *a little* progress on this — for instance, figuring out tricky ways in which she would be able to tell that some color, whatever it is, is *not* yellow, or *not* red. How? By

noting some salient and specific reaction that her brain would have only for yellow or only for red. But if you allow her even a little entry into her color space in this way, you should conclude that she can leverage her way to complete advance knowledge, because she doesn't just know the *salient* reactions, she knows them all.

Recall Julius and Ethel Rosenberg's Jell-O box, which they turned into an *M*-detector. Now imagine their surprise if an impostor were to show up with a "matching" piece that was not the original. "Impossible!" they cry. "Not impossible," says the impostor, "just difficult. I had *all the information* required to reconstruct an *M*-detector, and to make another thing with shape-property *M*." Mary had enough information (in the original case, if correctly imagined) to figure out just what her red-detectors and blue-detectors were, and hence to identify them in advance. Not the usual way of coming to learn about colors, but Mary is not your usual person.

I know that this will not satisfy many of Mary's philosophical fans, and that there is a lot more to be said, but — and this is my main point — the actual proving must go on in an arena far removed from Jackson's example, which is a classic provoker of Philosophers' Syndrome: mistaking a failure of imagination for an insight into necessity. Some of the philosophers who have dealt with the case of Mary may not care that they have imagined it wrong, since they have simply used it as a springboard into discussions that shed light on various independently interesting and important issues. I will not pursue those issues here, since I am interested in directly considering the conclusion that Jackson himself draws from his example: visual experiences have qualia that are "epiphenomenal."

The term "epiphenomena" is in common use today by both philosophers and psychologists (and other cognitive scientists). It is used with the presumption that its meaning is familiar and agreed upon, when in fact, philosophers and cognitive scientists use the term with *entirely* different meanings — a strange fact made even stranger to me by the fact that although I have pointed this out time and again, no one seems to care. Since "epiphenomenalism" often seems to be the last remaining safe haven for qualia, and since this appearance of safety is due entirely to the confusion between these two meanings, I must become a scold, and put those who use the term on the defensive.

According to the *Shorter Oxford English Dictionary*, the term "epiphenomenon" first appears in 1706 as a term in pathology, "a secondary appearance or symptom." The evolutionary biologist Thomas Huxley (1874) was probably the writer who extended the term to its current

use in psychology, where it means a *nonfunctional* property or by-product. Huxley used the term in his discussion of the evolution of consciousness and his claim that epiphenomenal properties (like the "whistle of the steam engine") could not be explained by natural selection.

Here is a clear instance of this use of the word:

> Why do people who are thinking hard bite their lips and tap their feet? Are these actions just epiphenomena that accompany the core processes of feeling and thinking or might they themselves be integral parts of these processes? [Zajonc and Markus, 1984, p. 74]

Notice that the authors mean to assert that these actions, while perfectly detectable, play no enabling role, no designed role. in the processes of feeling and thinking; they are nonfunctional. In the same spirit, the hum of the computer is epiphenomenal, as is your shadow when you make yourself a cup of tea. Epiphenomena are mere by-products, but as such they are products with lots of effects in the world: tapping your feet makes a recordable noise, and your shadow has its effects on photographic film, not to mention the slight cooling of the surfaces it spreads itself over.

The standard philosophical meaning is different: "x is epiphenomenal" means "x is an effect but itself has no effects in the physical world whatever." (See Broad, 1925, p. 118, for the definition that inaugurates, or at any rate establishes, the philosophical usage.) Are these meanings really so different? Yes, as different as the meanings of murder and death. The philosophical meaning is stronger: Anything that has no effects whatever in the physical world surely has no effects on the function of anything, but the converse doesn't follow, as the example from Zajonc and Markus makes obvious.

In fact, the philosophical meaning is too strong; it yields a concept of no utility whatsoever (Harman, 1990; Fox, 1989). Since x has no physical effects (according to this definition), no instrument can detect the presence of x directly or indirectly; the way the world goes is not modulated in the slightest by the presence or absence of x. How then, could there ever be any empirical reason to assert the presence of x? Suppose, for instance, that Otto insists that he (for one) has epiphenomenal qualia. Why does he say this? Not because they have some effect on him, somehow guiding him or alerting him as he makes his avowals. By the very definition of epiphenomena (in the philosophical sense), Otto's heartfelt avowals that he has epiphenomena could not

be evidence for himself or anyone else that he does have them, since he would be saying exactly the same thing even if he didn't have them. But perhaps Otto has some "internal" evidence?

Here there's a loophole, but not an attractive one. Epiphenomena, remember, are defined as having no effect in the *physical* world. If Otto wants to embrace out-and-out dualism, he can claim that his epiphenomenal qualia have no effects in the physical world, but do have effects in his (nonphysical) mental world (Broad, 1925, closed this loophole by definition, but it's free for the asking). For instance, they *cause* some *of his (nonphysical) beliefs*, such as his belief that he has epiphenomenal qualia. But this is just a temporary escape from embarrassment. For now on pain of contradiction, his beliefs, in turn, can have no effect in the physical world. If he suddenly lost his epiphenomenal qualia, he would no longer believe he had them, but he'd still go right on *saying* he did. He just wouldn't believe what he was saying! (Nor could he tell you that he didn't believe what he was saying, or do anything at all that revealed that he no longer believed what he was saying.) So the only way Otto could "justify" his belief in epiphenomena would be by retreating into a solipsistic world where there is only himself, his beliefs and his qualia, cut off from all effects in the world. Far from being a "safe" way of being a materialist and having your qualia too, this is at best a way of endorsing the most radical solipsism, by cutting off your mind — your beliefs and your experiences — from any commerce with the material world.

If qualia are epiphenomenal in the standard philosophical sense, their occurrence can't explain the way things happen (in the material world) since, by definition, things would happen exactly the same without them. There could not be an empirical reason, then, for believing in epiphenomena. Could there be another sort of reason for asserting their existence? What sort of reason? An *a priori* reason, presumably. But what? No one has ever offered one — good, bad, or indifferent — that I have seen. If someone wants to object that I am being a "verificationist" about these epiphenomena, I reply: Isn't everyone a verificationist about *this* sort of assertion? Consider, for instance, the hypothesis that there are fourteen epiphenomenal gremlins in each cylinder of an internal combustion engine. These gremlins have no mass, no energy, no physical properties; they do not make the engine run smoother or rougher, faster or slower. There is *and could be* no empirical evidence of their presence, and no empirical way in principle of distinguishing this hypothesis from its rivals: there are twelve or thirteen or fifteen . . . gremlins. By what principle does one defend one's

wholesale dismissal of such nonsense? A verificationist principle, or just plain common sense?

> Ah, but there's a difference! [says Otto.] There is no independent motivation for taking the hypothesis of these gremlins seriously. You just made them up on the spur of the moment. Qualia, in contrast, have been around for a long time, playing a major role in our conceptual scheme!

And what if some benighted people have been thinking for generations that gremlins made their cars go, and by now have been pushed back by the march of science into the forlorn claim that the gremlins are there, all right, but are epiphenomenal? Is it a mistake for us to dismiss their "hypothesis" out of hand? Whatever the principle is that we rely on when we give the back of our hand to such nonsense, it suffices to dismiss the doctrine that qualia are epiphenomenal in this philosophical sense. These are not views that deserve to be discussed with a straight face.

It's hard to believe that the philosophers who have recently described their views as epiphenomenalism can be making such a woebegone mistake. Are they, perhaps, just asserting that qualia are epiphenomenal in Huxley's sense? Qualia, on this reading, *are* physical effects and *have* physical effects; they just aren't functional. Any materialist should be happy to admit that this hypothesis is true — if we identify qualia with reactive dispositions, for instance. As we noted in the discussion of enjoyment, even though some bulges or biases in our quality spaces are functional — or used to be functional — others are just brute happenstance. Why don't I like broccoli? Probably for no reason at all; my negative reactive disposition is purely epiphenomenal, a by-product of my wiring with no significance. It has no function, but has plenty of effects. In any designed system, some properties are crucial while others are more or less revisable *ad lib*. Everything has to be some way or another, but often the ways don't matter. The gear shift lever on a car may have to be a certain length and a certain strength, but whether it is round or square or oval in cross section is an epiphenomenal property, in Huxley's sense. In the CADBLIND systems we imagined in chapter 10, the particular color-by-number coding scheme was epiphenomenal. We could "invert" it (by using negative numbers, or multiplying all the values by some constant) without making any *functional* difference to its information-processing prowess. Such an inversion might be undetectable to casual inspection, and might be undetectable *by the system*, but it would not be epiphenom-

enal in the philosophical sense. There would be lots of tiny voltage differences in the memory registers that held the different numbers, for instance.

If we think of all the properties of our nervous systems that enable us to see, hear, smell, taste, and touch things, we can divide them, roughly, into the properties that play truly crucial roles in mediating the information processing, and the epiphenomal properties that are more or less revisable *ad lib*, like the color-coding system in the CADBLIND system. When a philosopher surmises that qualia are epiphenomenal properties of brain states, this might mean that qualia could turn out to be local variations in the heat generated by neuronal metabolism. That cannot be what epiphenomenalists have in mind, can it? If it is, then qualia as epiphenomena are no challenge to materialism.

The time has come to put the burden of proof squarely on those who persist in using the term. The philosophical sense of the term is simply ridiculous; Huxley's sense is relatively clear and unproblematic — and irrelevant to the philosophical arguments. No other sense of the term has any currency. So if anyone claims to uphold a variety of epiphenomenalism, try to be polite, but ask: What *are* you talking about?

Notice, by the way, that this equivocation between two senses of "epiphenomenal" also infects the discussion of zombies. A philosopher's zombie, you will recall, is behaviorally indistinguishable from a normal human being, but is not conscious. There is nothing it is like to be a zombie; it just seems that way to observers (including itself, as we saw in the previous chapter). Now this can be given a strong or weak interpretation, depending on how we treat this indistinguishability to observers. If we were to declare that *in principle*, a zombie is indistinguishable from a conscious person, then we would be saying that genuine consciousness is epiphenomenal *in the ridiculous sense*. That is just silly. So we could say instead that consciousness might be epiphenomenal in the Huxley sense: although there was some way of distinguishing zombies from real people (who knows, maybe zombies have green brains), the difference doesn't show up as a functional difference to observers. Equivalently, human bodies with green brains don't harbor observers, while other human bodies do. On this hypothesis, we would be able in principle to distinguish the inhabited bodies from the uninhabited bodies by checking for brain color. This is also silly, of course, and dangerously silly, for it echoes the sort of utterly unmotivated prejudices that have denied full personhood to people on the basis of the color of their skin. It is time to recognize the idea of

the possibility of zombies for what it is: not a serious philosophical idea but a preposterous and ignoble relic of ancient prejudices. Maybe women aren't really conscious! Maybe Jews! What pernicious nonsense. As Shylock says, drawing our attention, quite properly, to "merely behavioral" criteria:

> Hath not a Jew eyes? Hath not a Jew hands, organs, dimensions, senses, affections, passions; fed with the same food, hurt with the same weapons, subject to the same diseases, heal'd by the same means, warm'd and cool'd by the same winter and summer, as a Christian is? If you prick us, do we not bleed? If you tickle us, do we not laugh? If you poison us, do we not die?

There is another way to address the possibility of zombies, and in some regards I think it is more satisfying. Are zombies possible? They're not just possible, they're actual. We're all zombies.[6] Nobody is conscious — not in the systematically mysterious way that supports such doctrines as epiphenomenalism! I can't prove that no such sort of consciousness exists. I also cannot prove that gremlins don't exist. The best I can do is show that there is no respectable motivation for believing in it.

6. GETTING BACK ON MY ROCKER

In chapter 2, section 2, I set up the task of explaining consciousness by recollecting an episode from my own conscious experience as I sat, rocking in my chair, looking out the window on a beautiful spring day. Let's return to that passage and see how the theory I have developed handles it. Here is the text:

> 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.

---

6. It would be an act of desperate intellectual dishonesty to quote this assertion out of context!

> 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. . . . 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? . . . It does seem impossible.

Since I have encouraged us all to be heterophenomenologists, I can hardly exempt myself, and I ought to be as content to be the subject as the practitioner, so here goes: I apply my own theory to myself. As heterophenomenologists, our task is to take this text, interpret it, and then relate the objects of the resulting heterophenomenological world of Dennett to the events going on in Dennett's brain at the time.

Since the text was produced some weeks or months after the events about which it speaks occurred, we can be sure that it has been abridged, not only by the author's deliberate editorial compressions, but also by the inexorable abridgment processes of memory over time. Had we probed earlier — had the author picked up a tape recorder while he sat rocking, and produced the text there and then — it would surely have been quite different. Not only richer in detail, and messier, but also, of course, reshaped and redirected by the author's own reactions to the very process of creating the text — listening to the actual sounds of his own words instead of musing silently. Speaking aloud, as every lecturer knows, often reveals implications (and particularly problems) in one's own message that elude one when one engages in silent soliloquy.

As it is, the text portrays a mere portion (and no doubt an idealized portion) of the contents of the author's consciousness. We must be careful, however, not to suppose that the "parts left out" in the given text were all "actually present" in something we might call the author's stream of consciousness. We must not make the mistake of supposing that there are facts — unrecoverable but actual facts — about just which contents were conscious and which were not at the time. And in particular, we should not suppose that when he looked out the window, he "took it all in" in one wonderful mental gulp — even though this is what his text portrays. It seemed to him, according to the text, as if

his mind — his visual field — were filled with intricate details of gold-green buds and wiggling branches, but although this is how it seemed, this was an illusion. No such "plenum" ever came into his mind; the plenum remained out in the world where it didn't have to be *represented*, but could just *be*. When we marvel, in those moments of heightened self-consciousness, at the glorious richness of our conscious experience, the richness we marvel at is actually the richness of the world outside, in all *its* ravishing detail. It does not "enter" our conscious minds, but is simply available.

What about all the branches and twigs rippling in unison? The branches outside on the tree didn't ripple, to be sure, since the rippling was due to the wrinkle in the windowpane, but that doesn't mean that all that rippling had to be happening in the author's mind or brain, just that it happened inboard of the windowpane that caused it. If someone had filmed the changing images on the author's retinas, they would have found the rippling there, just as in a movie, but that was no doubt where almost all the rippling stopped; what happened inboard of his retinas was just his recognition *that* there was, as he says in the text, a wonderful wave of synchronized ripples for him to experience. He saw the ripples, and he saw the extent of them, in just the way you would see all the Marilyns in the wallpaper. And since his retinas were provided with a steady dose of rippling, had he felt like sampling it further, there *would have been* more detail in the Multiple Drafts of which our text is all that remains.

There were many other details that the author could have focused on, but didn't. There are plenty of unrecoverable but genuine facts of the matter about which of these details got discriminated where and when by various systems in his brain, but the sum total of those facts doesn't settle such questions as which of these was he definitely, actually conscious of (but had forgotten by the time he produced his text), and which were definitely, actually in the "background" of his consciousness (though he didn't attend to them at the time). Our tendency to suppose that there *has* to be a fact of the matter to settle such questions is like the naïve reader's supposition that there *has* to be an answer to such questions as: Did Sherlock Holmes have eggs for breakfast on the day that Dr. Watson met him? Conan Doyle *might have* put that detail into the text, but he didn't, and since he didn't, there is simply no fact of the matter about whether those eggs belong *in the fictional world of Sherlock Holmes*. Even if Conan Doyle *thought* of Holmes eating eggs that morning, even if in an early draft Holmes *is represented in handwritten words* as eating eggs that morning, there is simply no

fact of the matter about whether in the fictional world of Sherlock Holmes, the world constituted from the published text we actually have, he had eggs for breakfast.

The text we have from Dennett was not "written in his brain" between the time in the rocking chair and the time it was typed into a file on a word processor. The attending he engaged in while rocking, and the concomitant rehearsal of those particulars that drew his attention, had the effect of fixing the contents of those particulars relatively securely "in memory" but this effect should not be viewed as storing a picture (or a sentence) or any other such salient representation. Rather, it should be thought of as just making a partially similar recurrence of the activity more likely, and that likely event is what happened, we may presume, on the occasion of the typing, driving the word-demons in his brain into the coalitions that yielded, for the first time, a string a sentences. Now some of what happened earlier, in the rocking chair, no doubt enlisted actual English words and phrases, and this prior collaboration between wordless contents and words no doubt facilitated the recovery of some of the very same English expressions when typing time came around.

Let's return to the heterophenomenological world of that text. What about the joy of which it speaks? ". . . the combination of sunny light, sunny Vivaldi violins, rippling branches — plus the pleasure I took in just thinking about it all. . . ." This could not be explained by the invocation of intrinsically pleasant qualia of sight, sound, and sheer thought. The idea that there are such qualia just distracts us from all possible paths of explanation, capturing our attention the way a wagging finger in front of a baby's eyes can capture its attention, getting us to stare numbly at the "intrinsic object" instead of casting about for a description of the underlying mechanisms *and* an explanation (ultimately an evolutionary explanation) of why the mechanisms do what they do.

The author's enjoyment is readily explainable by the fact that all visual experience is composed of the activities of neural circuits whose very activity is innately pleasing to us, not only because we simply like to become informed but because we like the particular ways we come to be informed. The fact that the look of sunlight-dappled spring buds should be something a human being likes is not surprising. The fact that some human beings also like looking at microscopic slides of bacteria and others like looking at photographs of airplane crashes is stranger, but the sublimations and perversions of desire grow from the same animal sources in the wiring of our nervous systems.

The author goes on to wonder how on earth "All this could indeed be just a combination of electrochemical happenings in my brain." As his wondering makes plain, it doesn't seem to be. Or in any event there was a moment when it occurred to him that it didn't seem to him to be just a combination of electrochemical happenings in his brain. But our subsequent chapters suggest a retort: Well, what do you think it would seem like if it were just a combination of electrochemical happenings in your brain?[7] Haven't we given ourselves grounds for concluding that with a brain organized the way ours is, this is just the sort of heterophenomenological world we would expect? Why shouldn't such combinations of electrochemical happenings in the brain have precisely the effects we set out to explain?

(The author speaks:) There is still one puzzle, however. How do *I* get to know all about this? How come *I* can tell you all about what was going on in my head? The answer to the puzzle is simple: *Because that is what I am.* Because a knower and reporter of such things in such terms is what is me. *My* existence is explained by the fact that there are these capacities in this body.

This idea, the idea of the Self as the Center of Narrative Gravity, is one we are finally ready to examine. It is certainly an idea whose time has come. Imagine my mixed emotions when I discovered that before I could get my version of it properly published in a book,[8] it had already been satirized in a novel, David Lodge's *Nice Work* (1988). It is apparently a hot theme among the deconstructionists:

> According to Robyn (or, more precisely, according to the writers who have influenced her thinking on these matters), there is no such thing as the "Self" on which capitalism and the classic novel are founded — that is to say, a finite, unique soul or essence that constitutes a person's identity; there is only a subject position in an infinite web of discourses — the discourses of power, sex, fam-

---

7. Cf. Lockwood (1989): "What would consciousness have felt like if it had felt like billions of tiny atoms wiggling in place?" (pp. 15–16)

8. I presented the main ideas in my reflections on Borges, in *The Mind's I* (Hofstadter and Dennett, 1981, pp. 348–352), and drew them together in a talk, "The Self as the Center of Narrative Gravity," presented at the Houston Symposium in 1983. While waiting for that symposium volume to appear, I published a somewhat truncated version of my talk in the *Times Literary Supplement*, Sept. 16–22, 1988, under the boring title — not mine — "Why everyone is a novelist." The original version, under the title "The Self as the Center of Narrative Gravity," is still forthcoming in F. Kessel, P. Cole, and D. Johnson, eds., *Self and Consciousness: Multiple Perspectives*, Hillsdale, NJ: Erlbaum.

> ily, science, religion, poetry, etc. And by the same token, there is no such thing as an author, that is to say, one who originates a work of fiction *ab nihilo*. . . . in the famous words of Jacques Derrida . . . "*il n'y a pas de hors-texte*", there is nothing outside the text. There are no origins, there is only production, and we produce our "selves" in language. Not "you *are what you eat*" but "you *are what you speak*," or, rather "you *are what speaks* you," is the axiomatic basis of Robyn's philosophy, which she would call, if required to give it a name, "semiotic materialism."

Semiotic materialism? Must *I* call it that? Aside from the allusions to capitalism and the classic novel, about which I have kept my counsel, this jocular passage is a fine parody of the view I'm about to present. (Like all parody, it exaggerates; I wouldn't say there is *nothing* outside the text. There are, for instance, all the bookcases, buildings, bodies, bacteria . . .)

Robyn and I think alike — and of course we are *both*, by our own accounts, fictional characters of a sort, though of a slightly different sort.