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.