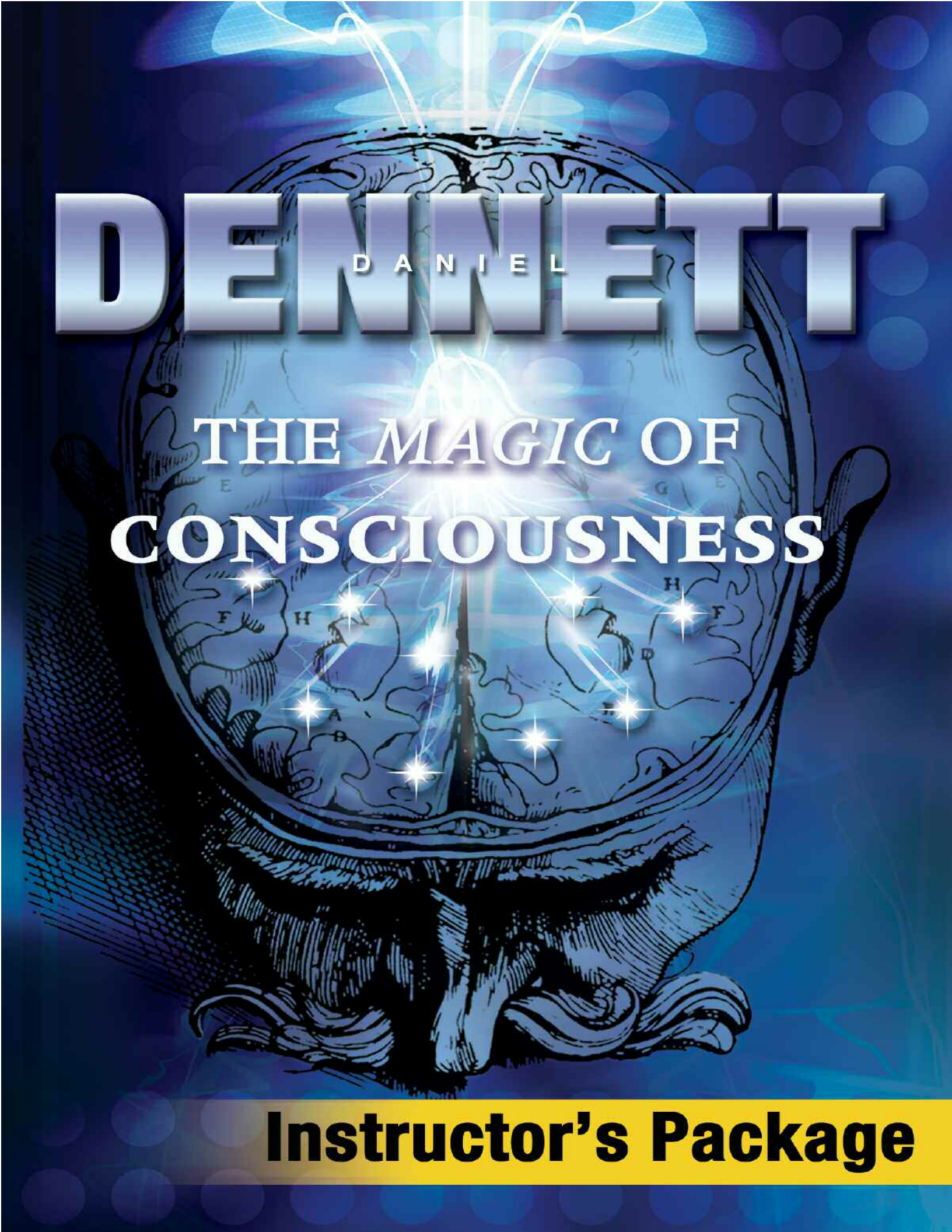

DANIEL
DENNETT
THE MAGIC OF
CONSCIOUSNESS
Instructor's Package

About This Guide

This guide is intended to assist in the use of the DVD *Daniel Dennett, Magic of Consciousness*.

The following pages provide an organizational schema for the DVD along with general notes for each section, key quotes from the DVD, and suggested discussion questions relevant to the section.

The program is divided into seven parts, each clearly distinguished by a section title during the program.

Contents

Seven-Part DVD

Articles by Daniel Dennett

1. The Problem of Consciousness

What should be so hard about understanding our own consciousness? It seems that there is nothing with which we're more familiar than the content of our own minds, nothing that we know better than our own thoughts. And how could this amazing phenomenon, our mental life, be produced by and situated inside our brain? This is one of the reasons that consciousness is such an interesting philosophical problem.

Consciousness is both the most familiar thing to all of us and one of the most mysterious. What could be more familiar to you than your own stream of consciousness? And yet, how on earth could it fit inside your brain? How on earth could what goes on in your brain actually account for what goes on [in your mind]?

Dennett gives us the metaphor of a "thought bubble" as a paradigm of how our own stream of consciousness feels to us, of what our mental life is like. It seems as if there is an entire world full of colors and sounds inside our heads, produced somehow by our brain and senses.

Then what's the problem? The problem is [that, as] vivid and brilliant as this representation of this man's consciousness is, it's metaphorical. *This is a large systematic metaphor for what's going on in his head.*

So the problem of consciousness, I would say, is if that's the metaphorical truth about what's going on in a man's head, what's the literal truth that makes the metaphor so good? What's actually happening between his ears that makes this such a brilliant metaphorical description of what's going on in the world?

So this is our challenge. How can it be that our vivid conscious experience of the world and of our own thoughts is produced by an organism composed of billions of tiny, unthinking, unconscious cells? How can we explain the phenomenon of consciousness so that it fits with the fact that it is a process performed by the trillions of cells, billions of neurons, that compose our brain?

Well, you might say, "What's the problem? Why is there any problem at all?" Well, you'll notice that Steinberg has represented the man as a sort of wonderful, pointillist collection of dots. And that's brilliant because it reminds us of what

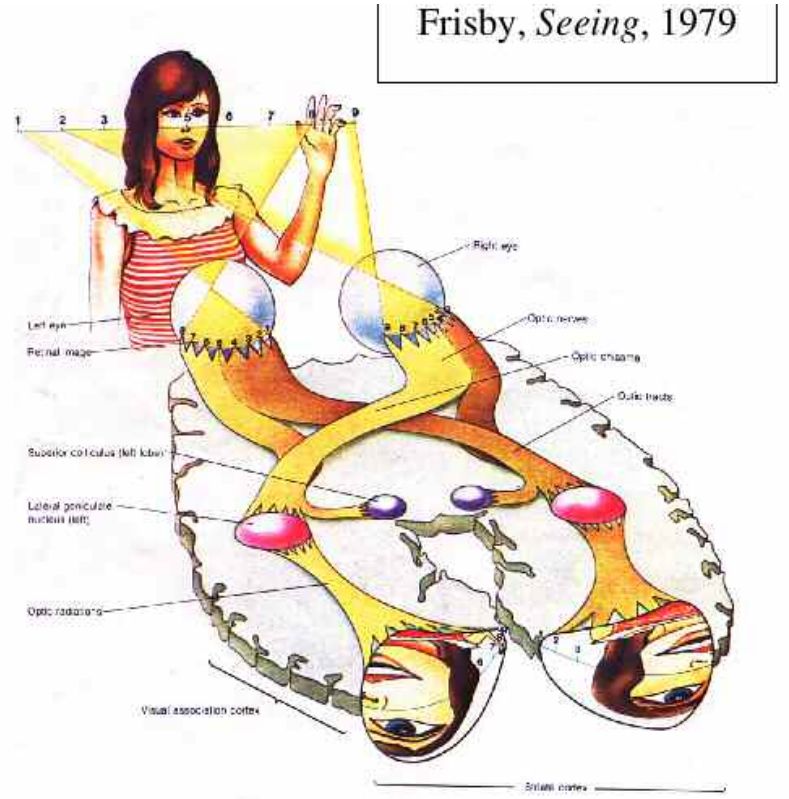
we are. What you are, what I am, is a huge collection of cells, about one hundred trillion at most recent count, a hundred trillion cells. That's what you are. Not a single one of those is conscious. Not a single one of those knows who Georges Braque is. Not a single one of those knows what Kilimanjaro is.

How can a collection of mindless, unconscious little robotic cells work together to create human consciousness as revealed in that beautiful metaphor? That's the problem of consciousness. And it's really quite severely puzzling.

The process of consciousness must happen somewhere in the brain, but finding the precise place or precise moment that our brain processes transform from unconscious information processing to conscious experience seems like an impossible task. There are plenty of areas of the brain that we could identify as performing different types of work, but it isn't obvious that any of these areas in particular are providing us, the subject of conscious experiences, with information of which we're conscious.

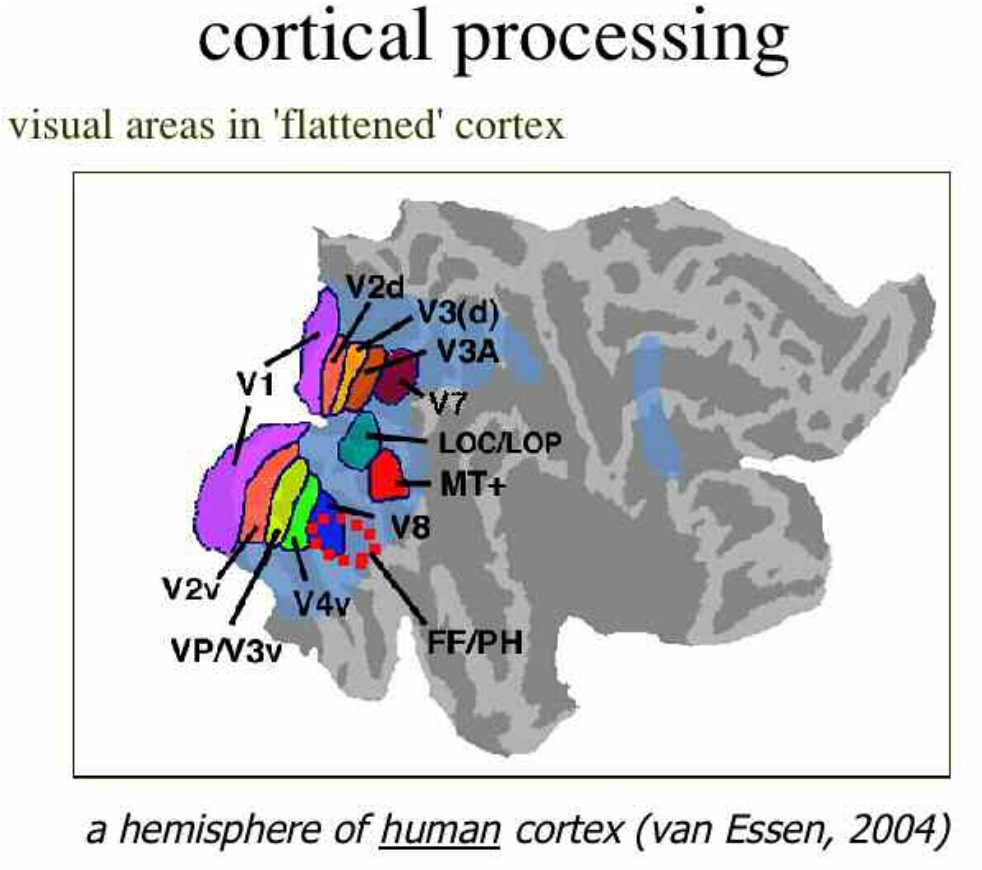
To see why [the process of consciousness is puzzling], let's just take a few simple steps. Now this is a vivid diagram from a textbook on vision by Frisby of some years ago. And what you can see here is just a little account of what happens so that you can become conscious of this woman standing in front of you. The light bounces off the woman and is focused by the lens. And these are the eyeballs, and an image is formed on the retina. And if you look in there with the right tools, you can see the image. It's a real image. It's upside-down, of course….

You see that the artist has drawn the lips as red. Of course, there's nothing red happening there. But there is stimulation happening in that area of the brain, which has the shape of a pair of lips, in fact.



Frisby, *Seeing*, 1979

But that's clearly not where the consciousness happens. Well, what happens next?…Well, cortical processing—here we see the cortex….And you'll see that there are other areas....And these are different visual areas in the brain which specialize to some degree….And they're all doing different parts of the job. But then what happens?



Then what happens, indeed? Then the magic of consciousness happens. Well, what on earth does that mean? This clearly isn't where we stop. We have to come to grips with what happens next.

Discussion Topic:

- What are some of the tasks that have to be performed somewhere in the brain for you to see, consciously, the woman with the red lips?

2. The Indian Rope Trick

In this section, Dennett begins to explain his theory of how a natural organ like the brain could produce something as extraordinary as consciousness. He draws an analogy between magic and consciousness, and it is quite fitting. To many people, consciousness does seem magical, that is, beyond human explanation. The brain, according to Dennett, is a sort of magician, but as he will explain, consciousness isn't "real" magic; the magic the brain performs is a sort of stage magic. Dennett quotes a passage from Lee Siegel's book on street magic in India, one that describes many people's reaction to his explanation of consciousness:

There's a passage which has become a sort of talisman for me, the passage that I love to quote:

> "I'm writing a book on magic," I explain, and I'm asked, "Real magic?" By *real magic* people mean miracles, thaumaturgical acts, and supernatural powers. "No," I answer: "Conjuring tricks, not real magic."
>
> *Real magic*, in other words, refers to the magic that is not real, while the magic that is real, that can actually be done, is *not real magic*.

And bingo, this is the story of my life, because the same thing is true of consciousness for many people. Real consciousness, in the eyes of some people, isn't something you could possibly explain. There are even scientists who think that almost by definition consciousness defies explanation. It is beyond human explanation. If you've explained anything, what you haven't explained is consciousness because they think consciousness is real magic.

"Real" magic, most all of us believe, doesn't exist in our world. If there is any sort of magic at all in the world, it is stage magic: impressive tricks performed by ordinary people with amazing skill. Although stage magic may not be "real" magic, it is the only sort of magic that exists, and we should keep in mind just how spectacular feats of stage magic can be. It takes a person with extraordinary skill, but not *supernatural* skill, to make us believe that he has made an airplane disappear, or sawed a woman in half. Stage magicians are incredibly talented individuals, and the impressiveness of stage magic should not be overlooked.

To help us grasp the important connection between "real" magic, stage magic, and consciousness, Dennett gives an explanation of how stage magic works, that is, from the philosophical point of view. He describes the Indian rope trick, in which a street magician supposedly tosses a rope into the air and then succeeds in climbing up the magically suspended rope. But this, according to Dennett, isn't the *real* Indian rope trick; the real trick involves the magi-

cian's assistant ascending the rope and disappearing into thin air, followed by the magician, who carves the assistant into pieces that fall to the ground. The magician descends, gathers the pieces of his assistant into a basket, and when the basket is opened the assistant jumps out—whole and unscathed.

Now, hands up those of you who've seen the trick performed. Ah, a miss, not a one of you, but what a trick that would be if it could be done. In fact, although you probably share my view that nobody has ever done the Indian rope trick and nobody ever could, millions of people believe that the trick has been performed. For over a hundred years, there have been urban legends, strong convictions held by people all over India and the rest of the world that the Indian rope trick has been performed. And you know how it goes.

"Well, I didn't see it myself, but I have an uncle who lives in Lahore, and it was his next-door neighbor who saw the trick performed in Bombay," something like that. And you might think, "Well, now, doesn't that show in a certain sense that the Indian rope trick has been performed?" Because after all, what is a magic trick? It's getting people to believe that you've done this amazing thing. You got somebody to believe you sawed the lady in half when you didn't.

This is a crucial point in analogy between magic and the mind. The job of the magician is to give his audience certain beliefs, to make them believe that he has performed a particular trick. The job of the brain, as we shall see later in section 5, is to make you, the biological organism that you are, believe certain things about the world that you need to believe to survive.

Dennett gives several examples of how a stage magician could plausibly convince his audience that he has performed a certain trick.

Well, now that you know what the Indian rope trick is, I'm going to tell you how to do it. All right, so here's a philosopher. He's going to explain the Indian rope trick. How to do the Indian rope trick, are you ready? Here's Method 1. First, gather an audience and claim that you're going to perform the Indian rope trick. This first step is actually very important, as you will see. Second, well, this is a step I'm not so clear about. It's not my department. Third, the audience, many of them anyway, exit, claiming to have seen the Indian rope trick.

You're probably not satisfied with that explanation so I'm going to give you another one. One, gather an audience and claim that you're going to perform the Indian rope trick, very important first step. Two, drug them all or hypnotize them. Three, plant the posthypnotic suggestion that you've done the trick complete with lots of details of what it looked like and let them wake up.

Now this is, in fact, the method that has been surmised by many people including many magicians who have tried to explain how so many people in India can be so sure the Indian rope trick has been done.

A stage magician somehow convinces his audience to believe that a great trick has been performed, one that could only be accomplished by "real" magic. None of these explanations Dennett gives seem satisfactory, but this *is* stage magic.

You probably think it's cheating. But, come on. We're talking about stage magic. What's cheating in stage magic? It's okay to use accomplices. It's okay to use wires and mirrors and smoke and distraction. What are the limits on cheating in stage magic?

Well, it's an interesting topic, but who cares? It's just show business. When the topic is consciousness, however, people often feel a little bit differently.

Discussion Topics:

- What is the difference between "real" magic and stage magic? What does it mean to say that stage magic is the only type of magic that *is* real?

- What is it about Dennett's explanation of how a magician performs a feat of stage magic that makes the trick look like cheating? Could there be any form of stage magic that wasn't a sort of cheating?

3. Stage Magic of the Mind

The first "magical" phenomenon of consciousness we discuss in section 3 is déjà vu. Déjà vu, of course, is the sensation that what you are currently experiencing is something you have experienced before; literally, "already seen."

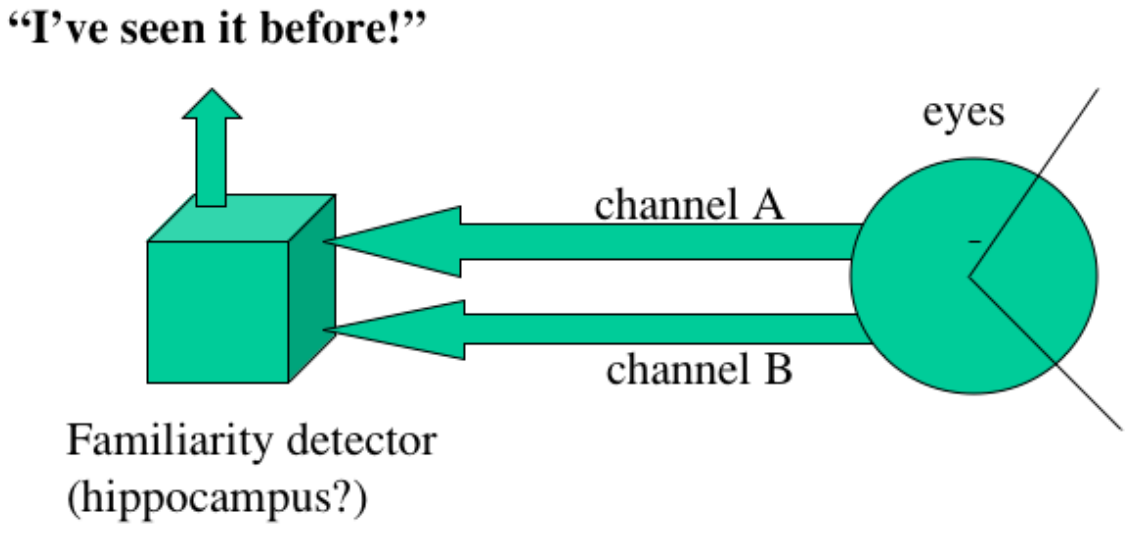
Many people think of déjà vu experiences as a sign of something pretty special, pretty amazing. Perhaps it's a glimpse that time is cyclical, and you can go round and round and relive our lives. Or maybe, it's transmigration of the soul. Or maybe, it's precognition. These are varying plausible or implausible quite dramatic hypotheses about the nature of déjà vu. We might call these [explanations of] déjà vu "magic hypotheses."

But since déjà vu is a phenomenon that occurs in the natural world to a natural, biological creature, the sensation of déjà vu must have a natural explanation. Dennett proceeds to give a scientifically plausible account of déjà vu inspired by the psychiatrist Pierre Janet, involving information pathways in the brain.

We imagine that the signal from the eye in the optic nerve and later parts of the [brain]—just suppose that it was split into two channels, channel A and B. And suppose that they were completely redundant. They're exactly the same. [What's in] channel B is just a duplicate of what's in channel A.

Just as some parts of the brain detect light and motion, other parts might function as something Dennett calls a "familiarity detector."

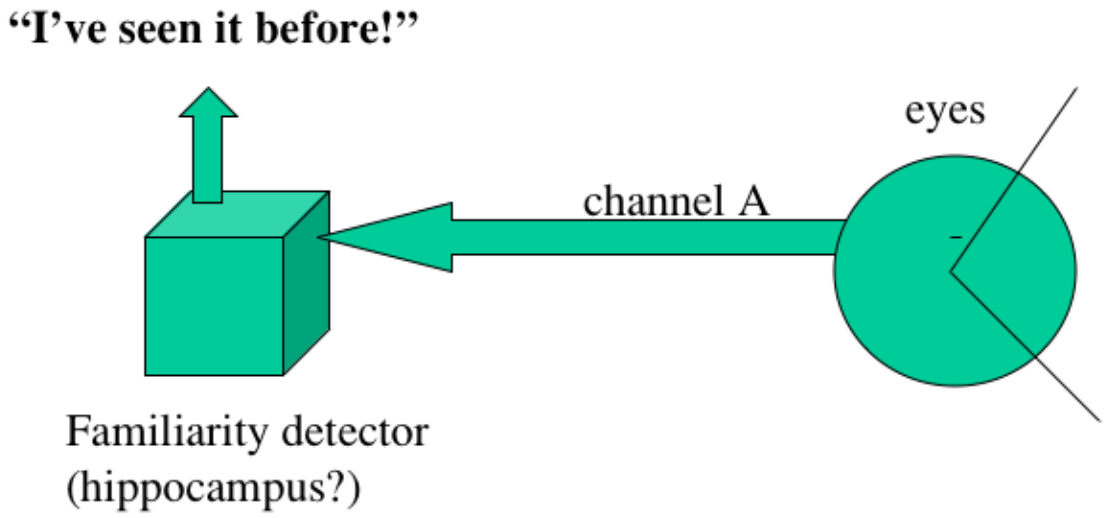
And then suppose that they arrive at a part of the brain that we're going to call the "familiarity detector"… a place in the brain which has the following properties: As new material comes in, it checks the material for novelty or familiarity. And if it's novel, it lets it through. And if it's familiar, it says, "Been there. Done that." It marks it as familiar. It sends a signal.



If there was a glitch in the timing between channel A and channel B, this might be enough to give a person the sensation that she had experienced the new thing she was seeing before.

Suppose that the signal in channel B got retarded in its passage from the eye to the detector by a few milliseconds, just long enough for channel A to go through and set up its signature, its footprint it's been there. It goes through [as] a novel event, and just a fraction of a second later, in comes the signal from channel B. And the system says, "Hey, I've seen it before!" Sure enough, ten milliseconds before, you saw it. But that's enough to send the "I've seen it before" signal on up into the system.

Dennett goes on to explain that we don't really need both channel A and channel B; all we need is one channel that can sometimes send the wrong signal up into the system.



All we have to suppose is that the familiarity detector on occasion, for who knows what reason, is spontaneously triggered. It just does a false positive. Suddenly it sends the "I've seen it before" signal spuriously up the line. And the rest, of course, is going to unfold just as it did in the first case.

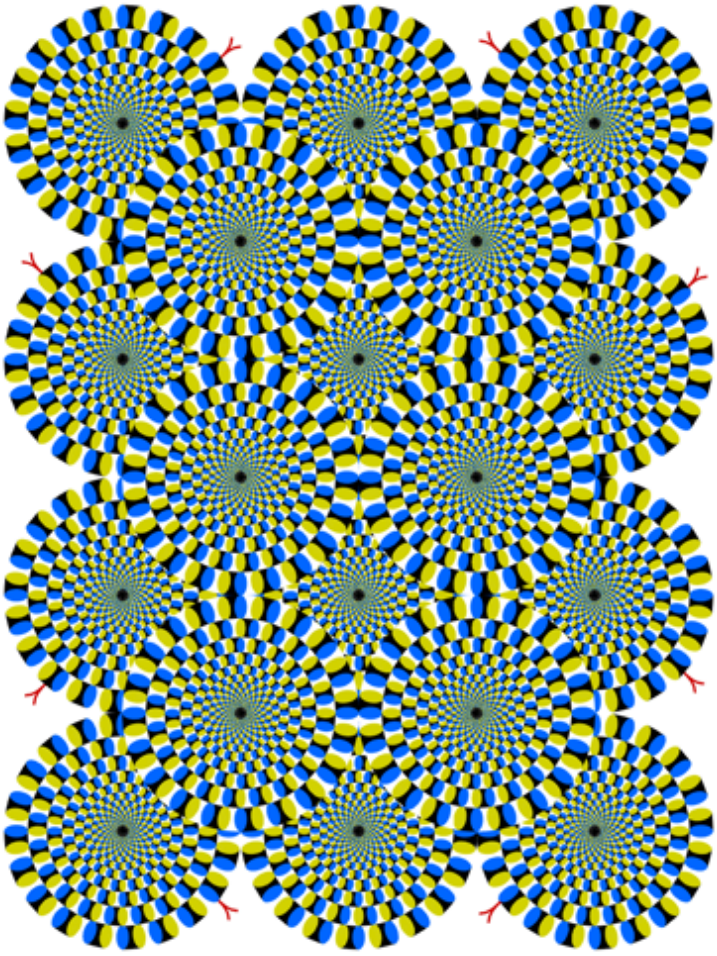
Now here's the amazing thing. It's a different hypothesis. You, when you have a déjà vu experience, you can't tell the difference between those two hypotheses. Either one of them could be the truth. And your own conscious experience of déjà vu doesn't give you any clue as to whether either of them might be the truth. This is something for third-person science, empirical investigation to explore.

Notice that in this second case, this is a philosopher telling you this. It's not that you've seen the thing before. It's that it [the malfunctioning detector mechanism] makes you think you've seen the thing before. That's all that has to happen for déjà vu to happen.

Either the first or second explanation could be true, and we have no way of knowing which is correct *from the inside*, from the first-person perspective. But there is a substantial difference between the two cases. In the first, the information first hits the detector in channel A, and so when it hits the detector in channel B the system *really has* "seen it before." In the second, the brain performs a bit of "stage magic"; it makes you *believe* you've seen it before

when all that has actually happened is that the "familiarity detector" *told* you you've seen it before.

There are other mental "illusions" that, though surprising, can be given similar explanations involving unexpected brain "signals." Dennett discusses cases of illusory motion in peripheral vision and afterimages. The first involves a series of circles that seem to move if one stares at the small black iris in the center. If the illusion is successful, the circles seem to turn against each other and rotate. But there are no rotating images in the brain, only systems of motion detection sending signals to other systems that *say* they detect motion.



"The shapes seem to be rotating!" Nothing is rotating in your brain. It just seems to you that the shapes are rotating.

Now there's nothing rotating up on the screen, nothing at all. But the shapes seem to be rotating, right? Nothing is rotating in your brain. If we were to look at the various visual areas where those parts are represented, we wouldn't see any rotation of any image. We would, however, see that regions in which motion detection occurs [send signals just as the regions for "familiarity detection" send signals] like "I've seen it before." [Signals] like, "something's moving"; those signals are being sent on up into the system. It just seems to you that the shapes are rotating.

The sensation of an afterimage can be even more puzzling, since surely there is nothing in the world that is red or striped, and yet there appears to be. Or, at least, our sensation of the flag afterimage is *like* the sensation we might have if we saw a translucent flag before us.

To explain how this might happen, Dennett brings in the tool of "maximally bland computationalism."

Ah, everybody got the American flag? There's no red stripes up on the screen. There's no red stripes on your retinal image. There's no red stripes in your visual cortex. And yet, it sure did seem as if there was a red stripe somewhere, didn't it?...I've gathered reports from people, and they say things like, "Well, the lowest short red stripe is intercepting the black cross." What? The lowest short red stripe is intercepting the black cross. So I want to know, "What are you talking about? What thing are you talking about?"

Well, you might say "something real" or "something red." You can refer to it. You can recall it, that red stripe. Where is it? It's not in your brain….There's nothing red happening in your brain when you see that image. There just seems to be something red happening in your brain. There's that philosopher's explanation again. Well, how does it happen?

I can't tell you. That's not my department. But I can give you some small clues. I want to describe a framework in which we could explain things like this, and I'm going to call it "maximally bland computationalism."

Discussion Topic:

- It seems to us that the stripes of the flag overlap with other things, for instance, the blackboard or the clock on the wall. It seems as if we could even point to the stripes. But certainly the stripes aren't "out there" in the world, so what might we mean when we say we see the stripes over the clock, or on the window?

4. Computationalism

To say that the brain is a computational system is just to say that it's a sort of computer. Computers, and brains, store information in what computer scientists call "registers." A register is just a name for whatever is holding information that is used in a larger, computational system. Dennett explains the fundamentals of what registers are and how they interact with or affect one another.

So we have a brain, and we want to understand that the brain is a computational system consisting of… trillions of registers. "Register" is a term from computer science. And a register is simply a memory location where you can store a number, a value. And it might be zero, or it might be one, or it might be 375, or it might be a million and three. A register is simply an address with a content, and the content is always a number; that is to say, it's always a magnitude of something.

So maximally bland computationalism says, what the brain is is a massively parallel, indeed, 3-D parallel collection of registers.

To say that the brain is "massively parallel" is to say that, unlike a typical home computer, in which just one thing happens at a time, creating a "serial" stream of computational events marching along in single file (but very, very swiftly), in a brain many of the registers can be computing at the same time, feeding new results to each other. This is how brains, whose parts are much slower acting than the parts of computers, can work faster than computers at many tasks. Why aren't laptops massively parallel? Because it is almost impossible to figure out how to program massively parallel machines. (If we ever do figure out how to program massively parallel computers, we should be able to make electronic brains that compute faster than organic brains.)

The information stored in a register affects the information stored in other registers, some close to that register, and some at a distance.

We suppose that the values are constantly shifting as a function of the values of other registers. And we explain all this just in terms of physics, just garden-variety, causal transactions between registers.

Now, I call this "maximally bland computationalism" because it makes no claims about the nature of the architecture. It's not a serial architecture, it's massively parallel. Is it asynchronous or synchronous? Well, presumably it's asynchronous. And the registers can be a neuron. You could consider that a register. Or you could consider it as made up of hundreds or thousands of registers.

Subcellular activity can be captured in this picture. Neuromodulator activity can be captured in this picture. Thus, if neuromodulators are being diffused through a part of the brain, there's a computational account of that. In fact, it's called "diffusive computation." And there are models which look at that.

The fact that the brain is a "wet" computer, that it is composed of organic material, does not affect the structure of how it functions. Neurons can serve as registers just as well as any system in your average computer can, and all this activity can be modeled as massively parallel processing. The brain is a sort of computer that takes in information from the world and distributes it throughout its subsystems in order to compute what would be the best thing for the person to do given that information.

The basic underlying idea is that it's the brain's job is to get the body that it resides in through life by computing the best thing to do next given the information that it's taking in from the world. That's what brains are for.... They're a control system, and hence, they're amenable to a computational analysis as long as we're suitably bland about what we mean by computational.

Discussion Topic:

- Given Dennett's abstract account of what the brain does for the body, that is, how it serves as a control system, what do students think this might mean for the possibility of artificial intelligence?

5. Reverse Engineering the Magic Show

Now that we've interpreted the brain as an immense computing machine, how should we explain what's happening when we see a flag in front of us that doesn't exist, or rotating circles that aren't actually spinning? Reverse engineering is the task of taking an existing machine or other artifact and figuring out how it works and why its parts are situated as they are. We need to reverse engineer the "magic show" of consciousness; we already know what seems to be going on, and now we must look behind the curtain of the brain and see how these phenomena might be produced. As before with déjà vu, these illusions depend on the brain sending signals that it *should* send when our eyes see a flag, or when circles in front of us are spinning.

Now we can go back to our red stripe and ask, "What's going on here? How can it seem to you that there's something red when there isn't?" Well, some of the computational events that would happen if you were seeing a real external red stripe are happening in you. And these events caused by this computation, some of the events that would happen when you have the conviction that you're seeing a red stripe. And that's all. That's all that has to happen. It's like déjà vu. You don't have to show the event twice as long as the conviction that you've seen it before is created. And as long as the conviction has some computational embodiment, you're home.

Dennett compares these mental illusions to the technique employed in a Belotto painting to represent people on a bridge. Your brain interprets the marks on the canvas as details of the people on the bridge, and from a certain distance it really seems as if the people have been painted in great detail. Your brain is telling you that the detail is there, in the painting, and so it looks to you to be quite detailed. Until you take a closer look.

So I…am noticing that there's, on this bridge in the sunlight, there's a whole lot of people moving across the bridge. And I wanted to see them up close. And so I began to get closer and closer and closer and closer and closer. And when I got up close, I actually yelled. I screamed, not a terrific scream, but a real yelp because what somewhat farther away had seemed wonderfully detailed, as I got closer, there was less detail than I thought there was. It was bizarre. The closer I got the less detail there was.

Now what was going on here?…The spots, the blobs of paint that he put so artfully on the canvas suggest people

with arms, and legs, and clothes, and belt buckles, and hats, plumes, carriages and all the rest. And the brain takes the suggestion.

The mechanisms by which the brain forms expectations is an issue outside of philosophy and within neuroscience, but we can speculate a bit about what is happening—what needs to happen—in the brain for us to form the judgments about the world that we form. To begin, we should work from the assumption that the brain would not expend more energy than it needs to to make us form the judgments that we form.

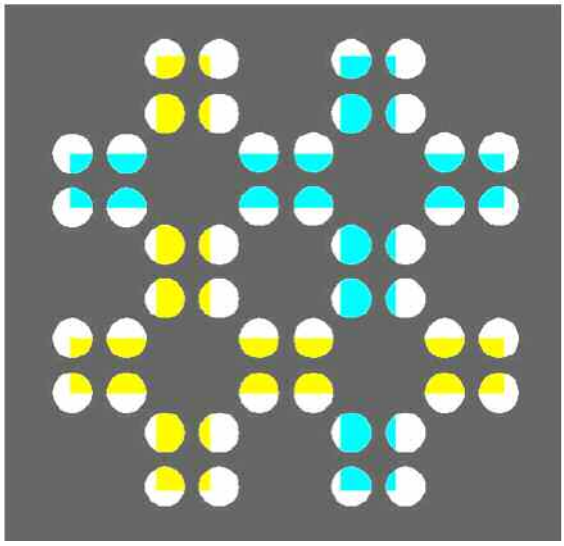
Well, now, how did the brain do this?... Did the brain, for instance, paint lots of little arms, and legs, and plumes, and hats, and buckles, and so forth, somewhere in one of those cortical areas where there's sort of an image of the scene and then look at them? I'm almost certain that's not the case. The brain didn't have to do that.

Well, then how did it do it?... I think it's a little bit like the posthypnotic suggestion.... All the brain has to do is create the judgment in a little part of the brain that is responsible for making judgments of that sort and then that judgment gets fed up into the system much like the déjà vu judgment where it can elaborate other judgments and play a causal role in causing other judgments and so forth. You don't have to do any painting in the brain for this to happen.

Dennett also gives the example of the Necker Cube, where the brain actually forms a representation of the lines where it assumes the lines would be. We could say here that the brain is "filling in" the missing detail. But this is not always how the brain operates—it doesn't always "fill in" the details, because they're not relevant to how we interpret an image. This is how the brain behaves in the case of the cube's color change.

Sometimes, as in the video clip of the moving vertical bars, there is an illusion that you are seeing moving edges "translucently" through the gray. This shows that we mustn't jump to conclusions, on the basis of introspection, about what the brain may do when it doesn't have to do it!

Rob van Lier, *Perception*



We learn not to trust our own conscious introspective experience for how these things really work. If you want to know how to explain the magic trick, you have to go backstage and see what's really going on.

Now, some of you may be thinking, "Wait a minute. This is all very interesting, but there's something fundamentally wrong with it." What I'm saying is that cognitive neuroscience can be seen as, in effect, reverse engineering the magic show, going to the staged magic show and showing you how the tricks are actually done, going backstage.

What needs to be explained, I've said, is what the audience thinks happened onstage. And this is, in fact, what I've called "heterophenomenonology." This is phenomenology of the third-person point of view. You gather lots of evidence about what the audience thinks is happening and then you have to explain why they think that.

Discussion Topic:

- What is the difference between the two kinds of "filling in" that occur in the case of the Necker cube seen floating in front of the disks or seen behind a screen with round holes in it? Do both kinds of filling in occur in normal perceptual circumstances?

6. The Dilemma of the Subject

We have said, so far, that the brain sends signals that tell us what's happening in the world—sometimes these signals accurately represent the world, and sometimes they don't. But one thing we seem to have left out of the explanation is the *us* to which the brain is sending these judgments. *We* are the ones having conscious experience, not our neurons, not our brains, and this is the part of consciousness we have left to explain.

In the standard view, consciousness is something like a movie unfolding in our minds, and we ourselves are the audience watching the show. Dennett calls this view "the Cartesian theater" picture of consciousness, in reference to philosopher Rene Descartes's view of the mind.

What am I doing talking about the audience?...One of the main themes in my work ever since 1991 in Consciousness Explained, *is that there's no such thing as the Cartesian theater. The Cartesian theater is my derisive negative term for the imaginary place in the brain where the inner witness, the audience, sits and enjoys the show of consciousness. And I've said, there is no such place.*

Dennett shows a picture of what the Cartesian theater could look like, with little men inside viewing the film of what's going on outside the body. But there is a major problem with this view—namely, that if there is someone watching the show, that someone must be conscious!

What's wrong with this is that if anything remotely like this were true, then we wouldn't have even started to explain consciousness because we would still have a conscious observer sitting there in the theater looking at the screen.

If a theory of consciousness is really going to explain consciousness, it can't do so by postulating an observer inside the conscious person, because then we'd need to explain how *that observer* is conscious: we'd have put another conscious observer inside the mind of that one, and so on forever. So if there's something that's already conscious in our theory of consciousness, we've only moved the problem back one step—we haven't really given a theory of consciousness at all!
We are going to need a theory of consciousness, then, that breaks consciousness down into parts that are not themselves conscious.

We're going to have to break up the Cartesian theater into parts that are not themselves conscious. The moral of the Cartesian theater [story] is that all the work done by the imaged homunculus in the Cartesian theater must be distributed around the various lesser agencies in the brain. And that means agencies that are not themselves conscious. Because if they were really conscious, then we've just re-created the homunculus problem, and we will not have made any progress at all.

Current neuroscience has shown that, not only is there no little observer within a person watching the "show" in the Cartesian theater, there is no point in the brain where all the sensory information and brain processes come together and become conscious.

The difficulty that most people have with this way of explaining consciousness it that, even though consciousness *must* be explained in this way (or else it would not qualify as an *explanation*), it seems to eliminate the *self* from the picture.

This creates a dilemma, though, because a lot of people are very unhappy with this. And I call this "the dilemma of the subject." What I've just said is, if you leave the subject (meaning I, ego, moi*)…in your theory, then you've not yet begun your theory of consciousness. You've just postponed the theory.*

However, the "dilemma of the subject" is seen by many philosophers as the main problem of consciousness; it has been labeled (by David Chalmers) the "hard problem" of consciousness. What happens to the self, the subject, when consciousness is broken down and explained in terms of smaller brain processes? Unfortunately for our intuitions about consciousness, this is the only way an explanation can proceed.

We have to agree that if you leave the subject in your theory you have not yet begun, so you have to get rid of the subject. And when you do this, the result has a certain scary, even disgusting feature. It's as if you've entered a factory, and there's all this humming machinery, and there's nobody home.

Dennett gives examples of two thinkers who have reacted strongly to this explanation of consciousness, Jerry Fodor and Bob Wright. Wright's reaction is reminiscent of our earlier remarks on "real" magic:

"Of course the problem here is with the claim that consciousness is 'identical' to physical brain states. The more Dennett and others try to explain to me what they mean by this, the more convinced I become that what they really mean is that consciousness doesn't exist." Now does that sound familiar? I hope so.

What Wright is saying is what people say about real magic. Real magic, in other words, refers to the magic that's not real while the magic that's real—that can actually be done—is not real magic. What Wright thinks is, if I'm saying that consciousness is a bunch of tricks in the brain, then what I'm really saying is that consciousness doesn't exist. There isn't any real magic.

"For Dennett, it is not a case of an emperor having no clothes, it's rather the clothes have no emperor." Yeah, that's right. Bingo. You've got to get rid of the emperor. If you still have the emperor in there, you don't have a theory of consciousness.

The same problem arises with a physical explanation of color. If we explain the phenomenon of color, say, red, by saying that it is produced by little red particles, then we haven't even begun to explain color. Because we then need to explain what makes these *particles* red. Although this seems completely obvious when we're talking about color, it is much harder to intuitively grasp this point when it comes to explaining consciousness.

Now that's a hard idea for people to get their heads around. But then, we knew, didn't we, that if you're going to explain consciousness, the result is going to have to be uncomfortable in some ways. It's going to have to be, as philosophers like to say, counterintuitive. Why? Because if there was an intuitive solution to the problems of consciousness, then we would have found it long ago. We've been working on this for several thousand years.

Discussion Topics:

- If we leave out the subject, what else has to go? What good would it be to have subjective colors, aromas, melodies in the brain if there was no subject to enjoy them? We know that there is no little band in the brain that plays the music we can have "running through our head." So what could possibly explain such phenomena?

- How does the "Cartesian theater" metaphor of the mind push the problem of consciousness back one step further?

- What are the similarities between "real" magic and "real" consciousness? What do students think of the criticism that Dennett's view has received? Is Dennett's theory a way of explaining consciousness, or does it deny the existence of consciousness?

7. The Magic of Consciousness

Dennett's view of consciousness is counterintuitive for many people, but a theory may be counterintuitive and *correct* at the same time. One of the intuitions that conflict with Dennett's view of consciousness is the "hard problem," mentioned in section 6. Dennett's reply to the "hard problem" of consciousness is that, in fact, philosophers are looking for something *too hard;* they overlook the possibility that what seems like one unified phenomenon of consciousness may break down into smaller, manageable pieces.

As an analogy to the puzzle of consciousness, Dennett gives the example of a magic trick performed by Ralph Hull that perplexed professional magicians for years: "the tuned deck."

It goes sort of like this. "Boys, I have a new trick." Now this is to his fellow magicians, right? "Boys, I have a new trick. It's called 'the tuned deck.' Here's my tuned deck. It is tuned. I listen to the vibrations, buzz, buzz, buzz, buzz. And by those vibrations I can tell exactly which card is here, is there, because of the different tuning of the vibrations. Here, pick a card, any card." The card is picked, goes back into the deck. There's some more shenanigans, some more buzz, buzz, and then the card is produced. That's it.

Now, Hull did this trick, hundreds of times, and nobody ever got it. He would sit with his sleeves rolled up and perform the trick for fellow magicians twenty, thirty times.

Late in his life, he gave the trick and his account of it to [John] Hilliard, his friend Hilliard, and Hilliard published it in his book. And here's a little bit of what Hull says. He says, "For years I've performed this effect and have shown it to magicians and amateurs by the hundred. And to the very best of my knowledge, not one of them ever figured out the secret.

"The boys have all looked for something too hard." Oh, thank you for saying "too hard." Now I'm going to tell you the secret of the tuned deck. Are you ready? The tuned deck, like many great magic tricks—the trick is over before you think it's even begun. In this case, the trick consists in its entirety in the title of the trick, "the tuned deck," moreover in one of the words in the title of the trick. Which word? No, not deck. Tuned, *no.* The. *I told you the trick is over before you even think it's begun.*

Hull's fellow magicians were searching for one big unified solution to the trick of the tuned deck, and therefore overlooked the simple answer: Hull was switching among many simple card tricks, and labeling the entire performance "*the* tuned deck."

Here is what Hull was doing. Remember how it starts, "Boys, I have a new trick. It's called 'the tuned deck.'" The trick is now over. . . . He does a standard card presentation trick that everybody there knows. The cards come back, and his fellow magicians think, "You know, couldn't he be doing a type-A trick?" . . .

So they're good magicians. They know how to prevent a type-A trick. So they . . . test their hypothesis that it's a type-A trick. He still does the trick. "Mm," they say, "could he be doing a type-B trick?" . . . So they do what they could do to prevent a type-B trick. He still does the trick. . . . No matter what hypothesis they test, he always does the trick. You like can't prevent him from doing the trick.

What's happened is, it was a type-A trick. Then, when they test that, he does a type-B trick. When they test B, he does a type-C trick. When they test that, he goes back, and he does a type-A trick. He realized he could always do one trick

or another, and he just did whichever trick they let him do.
And the reason they didn't tumble for it was the word the, *"the tuned deck." They were looking, as he said, for something too hard. They were looking for a hard problem, not a bunch of cheap tricks. In fact, all the tricks that they were doing were tricks that were quite familiar and in a certain way disappointing. And he hid all this with an elegant title.*

Just as magicians were tricked into thinking there was one great answer to the puzzle of the tuned deck, so philosophers are led to believe there must be one great answer to *the* problem of consciousness. But if we search for one great unified solution, instead of accepting that there may be smaller, manageable ways in which the brain produces consciousness, we might overlook the real answers to the problem of consciousness.

Now, I do want to suggest, but I don't claim to prove, that when David Chalmers talks about the "hard problem," he is innocently playing a trick on himself and others of exactly this sort. He's giving a name to a problem that doesn't even really exist. The problems of consciousness are how all of the various effects work. And once you've got an account of all those effects, that's what Chalmers calls the "easy problems," you're home. You've explained consciousness because there isn't any further problem, the hard problem. There just seems to be.

So again here's what Hull says about the tuned deck. He says, "Each time it's performed, the routine is such that one or more ideas in the back of the back of the spectator's head is exploded. Sooner or later, he will invariably give up any further attempt to solve the mystery."

Like many scientists and philosophers today, they just say, "It's mysterious. Give up. It's hopeless. We can't do it." Some of us think, "No, we can explain consciousness, but we have to be alert to the fact that many people want consciousness to be mysterious. They don't want it explained. They don't want it to be like stage magic. They want it to be like real magic." In other words, the kind of magic that isn't real.

So my conclusion is this: that the magic of consciousness, like stage magic, defies explanation only so long as we take it at face value. Once we appreciate all the nonmysterious ways in which the brain can create benign user illusions, we can begin to imagine how the brain creates consciousness.

At the end of the day, Dennett's message is this: to understand consciousness, we may need to reconsider our most fundamental presupposition about it—that it is one continuous, unified phenomenon. It may be that what we have taken to be *the* phenomenon of consciousness is really a collection of integrated individual, collaborative brain processes that *appear* to us as one great unified phenomenon. Explaining consciousness as a compilation of smaller processes may feel anticlimactic; we set out to investigate the Great Mystery of Consciousness, and our new task will be to examine ordinary brain activity. However, if the truth about the "magic" of consciousness is that it breaks down into nonmysterious subprocesses, we have arrived at our destination after all. What we must do now is trade in our fascination with the *magic* of consciousness for an appreciation of its *true* nature.

Discussion Topics:

- Hull's trick "the tuned deck" uses several smaller tricks to produce what seems to be one unified phenomenon. Discuss the analogy between the tuned deck and the problem of consciousness. How might we make similar claims about what the brain does to produce consciousness?

- What smaller brain processes (such as visual perception, information processing, etc.) might be analogous to the multiple tricks employed in the tuned deck?

Are We Explaining Consciousness Yet?

Abstract

Theorists are converging from quite different quarters on a version of the global neuronal workspace model of consciousness, but there are residual confusions to be dissolved. In particular, theorists must resist the temptation to see global accessibility as the *cause* of consciousness (as if consciousness were some other, further condition); rather, it *is* consciousness. A useful metaphor for keeping this elusive idea in focus is that consciousness is rather like fame in the brain. It is not a privileged medium of representation, or an added property some states have; it is the very mutual accessibility that gives some informational states the powers that come with a subject's consciousness of that information. Like fame, consciousness is not a momentary condition, or a purely dispositional state, but rather a matter of actual influence over time. Theorists who take on the task of accounting for the *aftermath* that is critical for consciousness often appear to be leaving out the Subject of consciousness, when in fact they are providing an analysis of the Subject, a necessary component in any serious theory of consciousness.

1. Clawing Our Way Towards Consensus

As the Decade of the Brain (declared by President Bush in 1990) comes to a close, we are beginning to discern how the human brain achieves consciousness. **Dehaene and Naccache** (this volume—all boldfaced citations below are to papers in this volume) see convergence coming from quite different quarters on a version of the global neuronal workspace model. There are still many differences of emphasis to negotiate, and, no doubt, some errors of detail to correct, but there is enough common ground to build on. I agree, and will attempt to re-articulate this emerging view in slightly different terms, emphasizing a few key points that are often resisted, in hopes of precipitating further consolidation. (On the eve of the Decade of the Brain, Baars (1988) had already described a "gathering consensus" in much the same terms: consciousness, he said, is accomplished by a "distributed society of specialists that is equipped with a working memory, called a *global workspace*, whose contents can be broadcast to the system as a whole" (p42). If, as **Jack and Shallice** point out, Baars' functional neuroanatomy has been superseded, this shows some of the progress we've made in the intervening years.)

A consensus may be emerging, but the seductiveness of the paths not taken is still potent, and part of my task here will be to diagnose some instances of backsliding and suggest therapeutic countermeasures. Of course those who still vehemently oppose this consensus will think it is I who needs therapy. These are difficult questions. Here is **Dehaene and Naccache**'s short summary of the global neuronal workspace model, to which I have attached some amplificatory notes on key terms, intended as friendly amendments to be elaborated in the rest of the paper:

> At any given time, many modular (1) cerebral networks are active in parallel and process information in an unconscious manner. An information (2) becomes conscious, however, if the neural population that represents it is mobilized by top-down (3) attentional amplification into a brain-scale state of coherent activity that involves many neurons distributed throughout the brain. The long distance connectivity of these "workplace neurons" can, when they are active for a minimal duration (4), make the information available to a variety of processes including perceptual categorization, long-term memorization, evaluation, and intentional action. We postulate that this global availability of information through the workplace is (5) what we subjectively experience as a conscious state. [from the ABSTRACT]

(1) Modularity comes in degrees and kinds; what is being stressed here is only that these are specialist networks with limited powers of information processing.

(2) There is no standard term for an event in the brain

that carries information or content on some topic (e.g., information about color at a retinal location, information about a phoneme heard, information about the familiarity or novelty of other information currently being carried, etc.). Whenever some specialist network or smaller structure makes a discrimination, fixes some element of content, "an information" in their sense comes into existence. "Signal," "content-fixation" (Dennett, 1991), "micro-taking" (Dennett and Kinsbourne, 1992), "wordless narrative" (Damasio, 1999), and "representation" (**Jack and Shallice**) are among the near-synonyms in use.

(3) We should be careful not to take the term "top-down" too literally. Since there is no single organizational summit to the brain, it means only that such attentional amplification is not just modulated "bottom-up" by features internal to the processing stream in which it rides, but also by *sideways* influences, from competitive, cooperative, collateral activities whose emergent net result is what we may lump together and call top-down influence. In an arena of opponent processes (as in a democracy) the "top" is distributed, not localized. Nevertheless, among the various competitive processes, there are important bifurcations or thresholds that can lead to strikingly different sequels, and it is these differences that best account for our pretheoretical intuitions about the difference between conscious and unconscious events in the mind. If we are careful, we can use "top-down" as an innocent allusion, exploiting a vivid fossil trace of a discarded Cartesian theory to mark the real differences that that theory misdescribed. (This will be elaborated in my discussion of **Jack and Shallice** below.)

(4) How long must this minimal duration be? Long enough to make the information available to a variety of processes—that's all. One should resist the temptation to imagine some *other* effect that needs to build up over time, because …

(5) The proposed consensual thesis is not that this global availability *causes* some further effect or a different sort altogether—igniting the glow of conscious qualia, gaining entrance to the Cartesian Theater, or something like that—but that it *is*, all by itself, a conscious state. This is the hardest part of the thesis to understand and embrace. In fact, some who favor the rest of the consensus balk at this point and want to suppose that global availability must somehow kindle some special effect over and above the merely computational or functional competences such global availability ensures. Those who harbor this hunch are surrendering just when victory is at hand, I will argue, for these "merely functional" competences are the very competences that consciousness was supposed to enable.

Here is where scientists have been tempted—or blackmailed—into defending unmistakably *philosophical* theses about consciousness, on both sides of the issue. Some have taken up the philosophical issues with relish, and others with reluctance and foreboding, with uneven results for both types. In this paper I will highlight a few of the points made and attempted, supporting some and criticizing others, but mainly trying to show how relatively minor decisions about word choice and emphasis can conspire to mislead the theoretician's imagination. Is there a "Hard Problem" (Chalmers, 1995, 1996) and if so what is it, and what could possibly count as progress towards solving it? Although I have staunchly defended—and will defend here again—the verdict that Chalmers' "Hard Problem" is a theorist's illusion (Dennett, 1996b, 1998), something inviting therapy, not a real problem to be solved with revolutionary new science, I view my task here to be dispelling confusion first, and taking sides second. Let us see, as clearly as we can, what the question is, and is not, before we declare any allegiances.

Dehaene and Naccache provide a good survey of the recent evidence in favor of this consensus, much of it analyzed in greater deal in the other papers in this volume, and I would first like to supplement their survey with a few anticipations drawn from farther afield. The central ideas are not new, though they have often been overlooked or underestimated. In 1959, the mathematician (and coiner of the term "artificial intelligence") John McCarthy, commenting on Oliver Selfridge's pioneering Pandemonium, the first model of a competitive, non-hierarchical computational architecture, clearly articulated the fundamental idea of the global workspace hypothesis:

> I would like to speak briefly about some of the advantages of the pandemonium model as an actual model of conscious behaviour. In observing

> a brain, one should make a distinction between that aspect of the behaviour which is available consciously, and those behaviours, no doubt equally important, but which proceed unconsciously. If one conceives of the brain as a pandemonium—a collection of demons—perhaps what is going on within the demons can be regarded as the unconscious part of thought, and what the demons are publicly shouting for each other to hear, as the conscious part of thought. (McCarthy, 1959, p147)

And in a classic paper, the psychologist Paul Rozin (1976), argued

> that specializations . . . form the building blocks for higher level intelligence. . . . At the time of their origin, these specializations are tightly wired into the functional system they were designed to serve and are thus inaccessible to other programs or systems of the brain. I suggest that in the course of evolution these programs become more *accessible* to other systems and, in the extreme, may rise to the level of consciousness and be applied over the full realm of behavior or mental function. (p246)

The key point, for both McCarthy and Rozin, is that it is the specialist demons' accessibility *to each other* (and not to some imagined higher Executive or central Ego) that could in principle explain the dramatic increases in cognitive competence that we associate with consciousness: the availability to deliberate reflection, the non-automaticity, in short, the open-mindedness that permits a conscious agent to consider anything in its purview in any way it chooses. This idea was also central to what I called the Multiple Drafts Model (Dennett, 1991), which was offered as an alternative to the traditional, and still popular, Cartesian Theater model, which supposes there is a place in the brain to which all the unconscious modules send their results for ultimate conscious appreciation by the Audience. The Multiple Drafts Model did not provide, however, a sufficiently vivid and imagination-friendly antidote to the Cartesian imagery we have all grown up with, so more recently I have proposed what I consider to be a more useful guiding metaphor: "fame in the brain" or "cerebral celebrity." (Dennett, 1994, 1996, 1998)

2. Competition for Clout

The basic idea is that consciousness is more like fame than television; it is *not* a special "medium of representation" in the brain into which content-bearing events must be transduced in order to become conscious. As **Kanwisher** aptly emphasizes: "the neural correlates of awareness of a given perceptual attribute are found in the very neural structure that perceptually analyzes that attribute." (ms, p6) Instead of switching media or going somewhere in order to become conscious, heretofore unconscious contents, staying right where they are, can achieve something *rather like* fame in competition with other fame-seeking (or just potentially fame-*finding*) contents. And, according to this view, that is what consciousness is.

Of course consciousness couldn't be *fame,* exactly, in the brain, since to be famous is to be a shared intentional object *in the conscious minds* of many folk, and although the brain is usefully seen as composed of hordes of demons (or *homunculi)*, if we were to imagine them to be *au courant* in the ways they would need to be to elevate some of their brethren to cerebral celebrity, we would be endowing these subhuman components with too much human psychology—and, of course, installing a patent infinite regress in the model as a theory of consciousness. The looming infinite regress can be stopped the way such threats are often happily stopped, not by abandoning the basic idea but by softening it. As long as your *homunculi* are more stupid and ignorant than the intelligent agent they compose, the nesting of homunculi within homunculi can be finite, bottoming out, eventually, with agents so unimpressive that they can be replaced by machines (Dennett, 1978). So consciousness is not so much fame, then, as political influence—a good slang term is *clout.* When processes compete for ongoing control of the body, the one with the greatest clout dominates the scene until a process with even greater clout displaces it. In some oligarchies, perhaps, the only way to have clout is to be *known by the King*, dispenser of all powers and privileges. Our brains are more democratic, indeed somewhat anarchic. In the brain there is no King, no Official Viewer of the State Television Program, no Cartesian Theater, but there are still plenty of quite sharp differences in political clout exercised by contents over time. In **Dehaene and Naccache**'s terms, this political difference is achieved by "reverberation" in a

"sustained amplification loop" (ms, p20), while the losing competitors soon fade into oblivion, unable to recruit enough specialist attention to achieve *self-sustaining* reverberation.

What a theory of consciousness needs to explain is how some relatively few contents become elevated to this political power, with all the ensuing *aftermath*, while most others evaporate into oblivion after doing their modest deeds in the ongoing projects of the brain. Why is this the task of a theory of consciousness? Because that is what conscious events do. They hang around, monopolizing time "in the limelight." We cannot settle for putting it that way, however. There is no literal searchlight of attention, so we need to explain away this seductive metaphor by explaining the functional powers of attention-*grabbing* without presupposing a single attention-*giving* source. This means we need to address two questions. Not just (1) How is this fame in the brain achieved? but also (2)—which I have called the Hard Question—And Then What Happens? (Dennett, 1991, p255). One may postulate activity in one neural structure or another as the necessary and sufficient condition for consciousness, but one must then take on the burden of the explaining why *that* activity ensures the political power of the events it involves—and this means taking a good hard look at how the relevant differences in competence might be enabled by changes in status in the brain.

Hurley (1998) makes a persuasive case for taking the Hard Question seriously in somewhat different terms: The Self (and its surrogates, the Cartesian *res cogitans*, the Kantian transcendental ego, among others) is not to be located by subtraction, by peeling off the various layers of perceptual and motor "interface" between Self and World. We must reject the traditional "sandwich" in which the Self is isolated from the outside world by layers of "input" and "output." On the contrary, the Self is large, concrete, and visible in the world, not just "distributed" in the brain but spread out into the world. Where we act and where we perceive is not funneled through a bottleneck, physical or metaphysical, in spite of the utility of such notions as "*point* of view."

As she notes, the very content of perception can change, while keeping input constant, by changes in output (p289).

This interpenetration of effects and contents can be fruitfully studied, and several avenues for future research are opened up by papers in this volume. What particularly impresses me about them is that the authors are all, in their various ways, more alert to the obligation to address the Hard Question than many previous theorists have been, and the result is a clearer, better-focused picture of consciousness in the brain, with no leftover ghosts lurking. If we set aside our *philosophical* doubts (settled or not) about consciousness as global fame or clout, we can explore in a relatively undistorted way the empirical questions regarding the mechanisms and pathways that are necessary, or just normal, for achieving this interesting functional status (we can call it a *Type-C* status, following **Jack and Shallice**, if we want to remind ourselves of what we are setting aside, while remaining noncommital). For example, **Parvizi and Damasio** claim that a midbrain panel of specialist proto-self evaluators accomplish a normal, but not necessary, evaluation process that amounts to a sort of triage, which can boost a content into reverberant fame or consign it to oblivion; these proto-self evaluators thereby tend to secure fame for those contents that are most relevant to current needs of the body. **Driver and Vuilleumier** concentrate on the "fate of extinguished stimuli" (ms, p18ff), exploring some of the ways that multiple competitions—e.g., as proposed by Desimone's and Duncan's (1995, 1995) Winner-Take-All model of multiple competition—leave not only single winners, but lots of quite powerful semi-finalists or also-rans, whose influences can be traced even when they don't achieve the canonical—indeed, operationalized—badge of fame: subsequent reportability (more on that, below). **Kanwisher** points out that sheer "activation strength" is no mark of consciousness until we see to what use that strength is put ("And then what happens?") and proposes that "the neural correlates of the *contents* of visual awareness are represented in the ventral pathway, whereas the neural correlates of more general-purpose *content-independent* processes associated with awareness (attention, binding, etc.) are found primarily in the dorsal pathway, which suggests (if I understand her claim rightly) that, just as in the wider world, whether or not you become famous can depend on what is going on *elsewhere* at the same time. **Jack and Shallice** propose a complementary balance between prefrontal cortex and anterior cingulate, a sort of high-road vs low-road dual path, with par-

ticular attention to the Hard Question: What can happen, what must happen, what may happen when Type-C processes occur, or put otherwise: what Type-C processes are necessary for, normal for, not necessary for. Particularly important are the ways in which successive winners dramatically alter the prospects (for fame, for influence) of their successors, creating nonce-structures that temporarily govern the competition. Such effects, described at the level of competition between "informations," can begin to explain how one (one agent, one subject) can "sculpt the response space" (Frith, 2000, discussed in **Jack and Shallice** [ms, p46]). This downstream capacity of one information to change the competitive context for whatever informations succeed it is indeed a fame-like competence, a hugely heightened influence that not only retrospectively distinguishes it from its competitors at the time but also, just as importantly, contributes to the creation of a relatively long-lasting Executive, not a place in the brain but a sort of political coalition that can be seen to *be in control* over the subsequent competitions for some period of time. Such differences in aftermath can be striking, perhaps never more so than those recently demonstrated effects that show, as **Dehaene and Naccache** note, "the impossibility for subjects [i.e., Executives] to strategically use the unconscious information," in such examples as Debner and Jacoby, 1994, and Smith and Merikle 1999, discussed in **Merikle et al.**

Consciousness, like fame, is not an *intrinsic* property, and not even just a *dispositional* property; it is a phenomenon that requires some actualization of the potential—and this is why you cannot make any progress on it until you address the Hard Question and look at the aftermath. Consider the following tale. Jim has written a remarkable first novel that has been enthusiastically read by some of the *cognoscenti*. His picture is all set to go on the cover of *Time* magazine, and Oprah has lined him up for her television show. A national book tour is planned and Hollywood has already expressed interest in his book. That's all true on Tuesday. Wednesday morning San Francisco is destroyed in an earthquake, and the world's attention can hold nothing else for a month. Is Jim famous? He would have been, if it weren't for that darn earthquake. Maybe next month, if things return to normal, he'll *become* famous for deeds done earlier. But fame eluded him this week, in spite of the fact that the *Time* magazine cover story had been typeset and sent to the printer, to be yanked at the last moment, and in spite of the fact that his name was already in *TV Guide* as Oprah's guest, and in spite of the fact that stacks of his novel could be found in the windows of most bookstores. All the *dispositional properties* normally sufficient for fame were in place, but their normal effects didn't get triggered, so no fame resulted. The same (I have held) is true of consciousness. The idea of some information being conscious for a few milliseconds, with none of the normal aftermath, is as covertly incoherent as the idea of somebody being famous for a few minutes, with none of the normal aftermath. Jim was potentially famous but didn't quite achieve fame; and he certainly didn't have any *other* property (an eerie glow, an aura of charisma, a threefold increase in "animal magnetism" or whatever) that distinguished him from the equally anonymous people around him. Real fame is not the *cause* of all the normal aftermath; it *is* the normal aftermath.

The same point needs to be appreciated about consciousness, for this is where theorists' imaginations are often led astray: it is a mistake to go looking for an *extra* will-of-the-wisp property of consciousness that might be enjoyed by some events in the brain in spite of their not enjoying the fruits of fame in the brain. Just such a quest is attempted by **Block**, who tries to isolate "phenomenality" as something distinct from fame ("global accessibility") but still worthy of being called a variety of consciousness. "Phenomenality is experience," he announces, but what does this mean? He recognizes that in order to keep phenomenality distinct from global accessibility, he needs to postulate, and find evidence for, what he calls "phenomenality without reflexivity"—experiences that you don't know you're having.

If we want to use brain imaging to find the neural correlates of phenomenality, we have to pin down the phenomenal side of the equation, and to do that we must make a decision on whether the subjects who say they don't see anything do or do not have phenomenal experiences.

But what then is left of the claim that phenomenality is experience? What is *experiential* (as contrasted with what?) about a discrimination that is not globally accessible? As the convolutions of Block's odyssey

reveal, there is always the simpler hypothesis to fend off: there is *potential* fame in the brain (analogous to the dispositional status of poor Jim, the novelist) and then there is fame in the brain, and these two categories suffice to handle the variety of phenomena we encounter. Fame in the brain is enough.

3. Is There Also a Hard *Problem*?

The most natural reaction in the world to this proposal is frank incredulity: it *seems* to be leaving out the most important element—the Subject! People are inclined to object: "There may indeed be fierce competition between 'informations' for political clout in the brain, but you have left out the First Person, who entertains the winners." The mistake behind this misbegotten objection is not noticing that the First Person has in fact already been incorporated into the multifarious further effects of all the political influence achievable in the competitions. Some theorists in the past have encouraged this mistake by simply stopping short of addressing the Hard Question. Damasio (1999) has addressed our two questions in terms of two intimately related problems: how the brain "generates the movie in the brain" and how the brain generates "the *appearance* of an owner and observer for the movie *within the movie,*" and has noted that some theorists, notably Penrose (1989) and Crick (1994), have made the tactical error of concentrating almost exclusively on the first of these problems, postponing the second problem indefinitely. Oddly enough, this tactic is reassuring to some observers, who are relieved to see that these models are not, apparently, denying the existence of the Subject but just not *yet* tackling that mystery. Better to postpone than to deny, it seems.

A model that, on the contrary, undertakes from the outset to address the Hard Question, assumes the obligation of accounting for the Subject in terms of "a collective dynamic phenomenon that does not require any supervision," as **Dehaene and Naccache** put it. This risks seeming to leave out the Subject, precisely because all the work the Subject would presumably have done, once it had enjoyed the show, has already been parceled out to various agencies in the brain, leaving the Subject with nothing to do. We haven't really solved the problem of consciousness until that Executive is itself broken down into subcomponents that are themselves *clearly* just unconscious underlaborers which themselves work (compete, interfere, dawdle . . .) without supervision. Contrary to appearances, then, those who work on answers to the Hard Question are not leaving consciousness *out*, they are explaining consciousness by leaving it *behind*. That is to say, the only way to explain consciousness is to move beyond consciousness, accounting for the effects consciousness has when it is achieved. It is hard to avoid the nagging feeling, however, that there must be something that such an approach leaves out, something that lies somehow in between the causes of consciousness and its effects.

Your body is made up of some trillions of cells, each one utterly ignorant of all the things *you* know. If we are to explain the conscious Subject, one way or another the transition from clueless cells to knowing organizations of cells must be made without any magic ingredients. This requirement presents theorists with what some see as a nasty dilemma (e.g., Andrew Brook, forthcoming). If you propose a theory of the knowing Subject that describes whatever it describes as like the workings of a vacant automated factory—not a Subject in sight—you will seem to many observers to have changed the subject or missed the point. On the other hand, if your theory still has tasks for a Subject to perform, still has a need for the Subject as Witness, then although you can be falsely comforted by the sense that there is still somebody at home in the brain, you have actually postponed the task of explaining what needs explaining. To me one of the most fascinating bifurcations in the intellectual world today is between those to whom it is obvious—*obvious*—that a theory that leaves out the Subject is thereby disqualified as a theory of consciousness (in Chalmers's terms, it evades the Hard Problem), and those to whom it is just as obvious that any theory that *doesn't* leave out the Subject is disqualified. I submit that the former have to be wrong, but they certainly don't lack for conviction, as these recent declarations eloquently attest:

> If, in short, there is a community of computers living in my head, there had also better be somebody who is in charge; and, by God, it had better be me. (Fodor, 1998, p207)

> Of course the problem here is with the claim that consciousness is 'identical' to physical brain states.

> The more Dennett et al. try to explain to me what they mean by this, the more convinced I become that what they really mean is that consciousness doesn't exist. (Wright, 2000, fn. 14, ch.21)
>
> *Daniel Dennett is the Devil*.... There is no internal witness, no central recognizer of meaning, and no self other than an abstract 'Center of Narrative Gravity' which is itself nothing but a convenient fiction.... For Dennett, it is not a case of the Emperor having no clothes. It is rather that the clothes have no Emperor. (Voorhees, 2000, pp55-56)

This is not just my problem; it confronts anybody attempting to construct and defend a properly naturalistic, materialistic theory of consciousness. Damasio is one who has attempted to solve this pedagogical (or perhaps diplomatic) problem by appearing to split the difference, writing eloquently about the Self, proclaiming that he is taking the Subject very seriously, even *restoring* the Subject to its rightful place in the theory of consciousness—while quietly dismantling the Self, breaking it into "proto-selves" and identifying these in functional, neuroanatomic terms as a network of brainstem nuclei (**Parvizi and Damasio**). This effort at winsome redescription, which I applaud, includes some artfully couched phrases that might easily be misread, however, as conceding too much to those who fear that the Subject is being overlooked. One passage in particular goes to the heart of current controversy. They disparage an earlier account that "dates from a time in which the phenomena of consciousness were conceptualized in exclusively behavioral, third-person terms. Little consideration was given to the cognitive, first-person description of the phenomena, that is, to the experience of the subject who is conscious." (ms, p2) Notice that they do *not* say that they are now adopting a first-person perspective; they say that they are now giving more consideration to the "first-person *description*" that subjects give. In fact, they are strictly adhering to the canons and assumptions of what I have called *heterophenomenology*, which is specifically designed to be a *third-person* approach to consciousness (Dennett, 1991, ch 4, "A Method for Phenomenology," p98). How does one take subjectivity seriously from a third-person perspective? By taking the *reports* of subjects seriously as reports of their subjective experience. This practice does not limit us to the study of human subjectivity; as numerous authors have noted, non-verbal animals can be put into circumstances in which some of their behavior can be interpreted, as Weiskrantz (1998) has put it, as "commentaries," and **Kanwisher** points out that in Newsome's experiments, for instance, the monkey's behavior is "a reasonable proxy for such a report." (ms, p4)

It has always been good practice for scientists to put themselves in their own experimental apparatus as informal subjects, to confirm their hunches about what it feels like, and to check for any overlooked or underestimated features of the circumstances that could interfere with their interpretations of their experiments. (**Kanwisher** gives a fine example of this, inviting the reader into the role of the subject in rapid serial visual display [RSVP], and noting from the inside, as it were, the strangeness of the forced choice task: you find yourself thinking that "tiger" would be as good a word as any, etc. [ms, p11]) But scientists have always recognized the need to confirm the insights they have gained from self-administered pilot studies by conducting properly controlled experiments with naive subjects. As long as this obligation is met, whatever insights one may garner from "first-person" investigations fall happily into place in "third-person" heterophenomenology. Purported discoveries that cannot meet this obligation may inspire, guide, motivate, illuminate one's scientific theory, but *they* are not data—the beliefs of subjects about them are the data. Thus if some phenomenologist becomes convinced by her own (first-)personal experience, however encountered, transformed, reflected upon, of the existence of a feature of consciousness in need of explanation and accommodation within her theory, her conviction that this is so is itself a fine datum in need of explanation, by her or by others, but the truth of her conviction must not be presupposed by science. There is no such thing as first-person science, so if you want to have a *science* of consciousness, it will have to be a third-person science of consciousness, and none the worse for it, as the many results discussed in this volume show.

Since there has been wholesale misreading of this moral in the controversies raging about the "first person point of view," let me take this opportunity to point out that every study reported in every article in this volume has been conducted according to the tenets of het-

erophenomenology. Are the researchers represented here needlessly tying their own hands? Are there other, deeper ways of studying consciousness scientifically? This has recently been claimed by Petitot, Varela, Pachoud and Roy (1999), who envision a "naturalized phenomenology" that somehow goes beyond heterophenomenology and derives something from a first-person point of view that cannot be incorporated in the manner followed here, but while their anthology includes some very interesting work, it is not clear that any of it finds a mode of scientific investigation that in any way even purports to transcend this third-person obligation. The one essay that makes such a claim specifically, Thompson, Noë and Pessoa's essay on perceptual completion or "filling in" (cf. Pessoa, Thompson and Noë, 1998) corrects some errors in my heterophenomenological treatment of the same phenomena, but is itself a worthy piece of heterophenomenology, in spite of the authors declarations to the contrary (see Dennett, 1998b, and their reply, same issue). Chalmers (1999) has made the same unsupported claim:

I also take it that first-person data can't be expressed wholly in terms of third-person data about brain processes *and the like* [my italics]. . . . .That's to say, no purely third-person description of brain processes *and behavior* [my italics] will express precisely the data we want to explain, though it may play a central role in the explanation. So 'as data,' the first-person data are irreducible to third-person data. (p8)

This swift passage manages to overlook the prospects of heterophenomenology altogether. Heterophenomenology is explicitly not a first-person methodology (as its name makes clear) but it is also not directly about "brain processes and the like"; it is a reasoned, objective extrapolation from patterns discernible in the behavior of subjects, including especially their text-producing or communicative behavior, and as such it is *about* precisely the higher-level dispositions, both cognitive and emotional, that convince us that our fellow human beings are conscious. By sliding from the first italicized phrase to the second (in the quotation above), Chalmers executes a (perhaps unintended) sleight-of-hand, whisking heterophenomenology off the stage without a hearing. His conclusion is a non sequitur. He has not shown that first-person data are irreducible to third-person data because he has not even considered the only serious attempt to show *how* first-person data can be "reduced" to third-person data (though I wouldn't use that term).

The third-person approach is not antithetical to, or eager to ignore, the subjective nuances of experience; it simply insists on anchoring those subjective nuances to *something*—anything, really—that can be detected and confirmed in replicable experiments. For instance, **Merikle et al.**, having adopted the position that "with subjective measures, awareness is assessed on the basis of the observer's self-reports," note that one of the assumptions of this approach is that "information perceived with awareness enables a perceiver to act on the world and to produce effects on the world." As contrasted to what? As contrasted to a view, such as that of Searle (1992) and Chalmers (1996), that concludes that consciousness *might* have no such enabling role—since a "zombie" might be able to do everything a conscious person does, passing every test, reporting every effect, without being conscious. One of the inescapable implications of heterophenomenology, or of any third-person approach to subjectivity, is that one must dismiss as a chimera the prospect of a philosopher's zombie, a being that is behaviorally, objectively indistinguishable from a conscious person but not conscious. (For a survey of this unfortunate topic, see *Journal of Consciousness Studies*, 2, 1995, "Zombie Earth: a symposium," including short pieces by many authors.)

I find that some people are cured of their attraction for this chimera by the observation that all the functional distinctions described in the essays in this volume would be exhibited by philosophers' zombies. The only difference between zombies and regular folks, according to those who take the distinction seriously, is that zombies have streams of *unconsciousness* where the normals have streams of *consciousness!* Consider, in this regard, the word-stem completion task of Debner and Jacoby (1994) discussed by **Merikle et al.** If subjects are instructed to complete a word stem with a word other than the word briefly presented as a prime (and then masked), they can follow this instruction only if they are aware of the priming word; they actually favor the priming word as a completion if it is presented so briefly that they are not aware of it. Zombies would exhibit the same effect, of course—being able to follow the exclusion policy only in those instances in which the priming word made it through the competition into their streams of *un*-consciousness.

4. But What About "Qualia"?

As **Dehaene and Naccache** note,

> [T]he flux of neuronal workspace states associated with a perceptual experience is vastly beyond accurate verbal description or long-term memory storage. Furthermore, although the major organization of this repertoire is shared by all members of the species, its details result from a developmental process of epigenesis and are therefore specific to each individual. Thus the contents of perceptual awareness are complex, dynamic, multi-faceted neural states that cannot be memorized or transmitted to others in their entirety. These biological properties seem potentially capable of substantiating philosophers' intuitions about the "qualia" of conscious experience, although considerable neuroscientific research will be needed before they are thoroughly understood. (ms, p34-5)

It is this informational superabundance, also noted by Damasio (1999, see esp p93), that has lured philosophers into a definitional trap. As one sets out to answer the Hard Question ("And then what happens?"), one can be sure that no practical, finite set of answers will exhaust the richness of effects and potential effects. The subtle individual differences wrought by epigenesis and a thousand chance encounters creates a unique manifold of functional (including *dys*functional) dispositions that outruns any short catalog of effects. These dispositions may be dramatic—ever since that yellow car crashed into her, one shade of yellow sets off her neuromodulator alarm floods (Dennett, 1991)—or minuscule—an ever so slight relaxation evoked by a nostalgic whiff of childhood comfort food. So one will always be "leaving something out." If one dubs this inevitable residue *qualia*, then qualia are guaranteed to exist, but they are just more of the same, dispositional properties that have not yet been entered in the catalog (perhaps because they are the most subtle, least amenable to approximate definition). Alternatively, if one defines *qualia* as whatever is neither the downstream effects of experiences (reactions to particular colors, verbal reports, effects on memory . . .) nor the upstream causal progenitors of experiences (activity in one cortical region or another), then qualia are, by definitional fiat, *intrinsic properties* of experiences considered in isolation from all their causes and effects, logically independent of all dispositional properties. Defined thus, they are logically guaranteed to elude all broad functional analysis—but it's an empty victory, since there is no reason to believe such properties exist! To see this, compare the qualia of experience to the value of money. Some naive Americans cannot get it out of their heads that dollars, unlike francs and marks and yen, have *intrinsic value* ("How much is that in *real* money?"). They are quite content to "reduce" the value of other currencies in dispositional terms to their exchange rate with dollars (or goods and services), but they have a hunch that dollars are different. Every dollar, they declare, has something logically independent of its functionalistic exchange powers, which we might call its *vis*. So defined, the *vis* of each dollar is guaranteed to elude the theories of economists forever, but we have no reason to believe in it—aside from their heartfelt hunches, which can be explained without being honored. It is just such an account of philosophers' intuitions that **Dehaene and Naccache** propose.

It is unfortunate that the term *qualia* has been adopted—in spite of my warnings (1988, 1991, 1994b)—by some cognitive neuroscientists who have been unwilling or unable to believe that philosophers intend that term to occupy a peculiar logical role in arguments about functionalism that cognitive neuroscience *could not* resolve. A review of recent history (drawn, with revisions, from Dennett, forthcoming) will perhaps clarify this source of confusion and return us to the real issues.

Functionalism is the idea enshrined in the old proverb: handsome is as handsome does. Matter matters only because of what matter can do. Functionalism in this broadest sense is so ubiquitous in science that it is tantamount to a reigning presumption of all of science. And since science is always looking for simplifications, looking for the greatest generality it can muster, functionalism in practice has a bias in favor of minimalism, of saying that less matters than one might have thought. The law of gravity says that it doesn't matter what stuff a thing is made of—only its mass matters (and its density, except in a vacuum). The trajectory of cannonballs of equal mass and density is not affected by whether they are made of iron, copper or gold. It might have mattered, one imagines, but in fact it doesn't. And wings don't have to have feathers on them in order to power

flight, and eyes don't have to be blue or brown in order to see. Every eye has many more properties than are needed for sight, and it is science's job to find the maximally general, maximally non-committal—hence minimal—characterization of whatever power or capacity is under consideration. Not surprisingly, then, many of the disputes in normal science concern the issue of whether or not one school of thought has reached too far in its quest for generality.

Since the earliest days of cognitive science, there has been a particularly bold brand of functionalistic minimalism in contention, the idea that just as a heart is basically a pump, and could in principle be made of anything so long as it did the requisite pumping without damaging the blood, so a mind is fundamentally a control system, implemented in fact by the organic brain, but anything else that could *compute the same control functions* would serve as well. The actual matter of the brain—the chemistry of synapses, the role of calcium in the depolarization of nerve fibers, and so forth—is roughly as irrelevant as the chemical composition of those cannonballs. According to this tempting proposal, even the underlying micro-architecture of the brain's connections can be ignored for many purposes, at least for the time being, since it has been proven by computer scientists that any function that can be computed by one specific computational architecture can also be computed (perhaps much less efficiently) by another architecture. If all that matters is the computation, we can ignore the brain's wiring diagram, and its chemistry, and just worry about the "software" that runs on it. In short—and now we arrive at the provocative version that has caused so much misunderstanding—in principle you could replace your wet, organic brain with a bunch of silicon chips and wires and go right on thinking (and being conscious, and so forth).

This bold vision, computationalism or "strong AI" (Searle, 1980), is composed of two parts: the broad creed of functionalism—handsome is as handsome does—and a specific set of minimalist empirical wagers: neuroanatomy doesn't matter; chemistry doesn't matter. This second theme excused many would-be cognitive scientists from educating themselves in these fields, for the same reason that economists are excused from knowing anything about the metallurgy of coinage, or the chemistry of the ink and paper used in bills of sale. This has been a good idea in many ways, but for fairly obvious reasons, it has not been a politically astute ideology, since it has threatened to relegate those scientists who devote their lives to functional neuroanatomy and neurochemistry, for instance, to relatively minor roles as electricians and plumbers in the grand project of explaining consciousness. Resenting this proposed demotion, they have fought back vigorously. The recent history of neuroscience can be seen as a series of triumphs for the lovers of detail. Yes, the specific geometry of the connectivity matters; yes, the location of specific neuromodulators and their effects matter; yes, the architecture matters; yes, the fine temporal rhythms of the spiking patterns matter, and so on. Many of the fond hopes of opportunistic minimalists have been dashed: they had hoped they could leave out various things, and they have learned that no, if you leave out *x*, or *y*, or *z*, you can't explain how the mind works.

This has left the mistaken impression in some quarters that the underlying idea of functionalism has been taking its lumps. Far from it. On the contrary, the reasons for accepting these new claims are precisely the reasons of functionalism. Neurochemistry matters because—and *only* because—we have discovered that the many different neuromodulators and other chemical messengers that diffuse through the brain have *functional roles* that make important differences. What those molecules *do* turns out to be important to the *computational* roles played by the neurons, so we have to pay attention to them after all.

This correction of over-optimistic minimalism has nothing to do with philosophers' imagined *qualia*. Some neuroscientists have thus muddied the waters by befriending qualia, confident that this was a term for the sort of functionally characterizable complication that confounds oversimplified versions of computationalism. (Others have thought that when philosophers were comparing zombies with conscious people, they were noting the importance of emotional state, or neuromodulator imbalance.) I have spent more time than I would like explaining to various scientists that their controversies and the philosophers' controversies are not translations of each other as they had thought but false friends, mutually irrelevant to each other. The principle of charity continues to bedevil this issue, however,

and many scientists generously persist in refusing to believe that philosophers can be making a fuss about such a narrow and fantastical division of opinion. Meanwhile, some philosophers have misappropriated those same controversies within cognitive science to support their claim that the tide is turning against functionalism, in favor of qualia, in favor of the irreducibility of the "first-person point of view," and so forth. This widespread conviction is an artifact of interdisciplinary miscommunication and nothing else. A particularly vivid exposure of the miscommunication can be found in the critics' discussion of Humphrey (2000). In his rejoinder Humphrey says:

> I took it for granted that everyone would recognise that my account of sensations was indeed meant to be a functional one through and through—so much so that I actually deleted the following sentences from an earlier draft of the paper, believing them redundant: "Thus [with this account] we are well on our way to doing the very thing it *seemed* we would not be able to do, namely giving the mind term of the identity, the phantasm, a *functional description*—even if a rather unexpected and peculiar one. And, as we have already seen, once we have a functional description we're home and dry, because the same description can quite well fit a brain state."

But perhaps I should not be amazed. Functionalism is a wonderfully—even absurdly—bold hypothesis, about which few of us are entirely comfortable.

5. Conclusion

A neuroscientific theory of consciousness must be a theory of the Subject of consciousness, one that analyzes this imagined central Executive into component parts, none of which can itself be a proper Subject. The apparent properties of consciousness that only make sense as *features enjoyed by the Subject* must thus also be decomposed and distributed, and this inevitably creates a pressure on the imagination of the theorist. No sooner do such properties get functionalistically analyzed into complex dispositional traits distributed in space and time in the brain, than their ghosts come knocking on the door, demanding entrance disguised as *qualia,* or *phenomenality* or *the imaginable difference between us and zombies*. One of the hardest tasks thus facing those who would explain consciousness is recognizing when some feature has *already* been explained (in sketch, in outline) and hence does not need to be explained again.

References

Brook, Andrew, forthcoming, "Judgments and Drafts Eight Years Later," in D. Ross and A. Brook, eds., *Dennett's Philosophy: A Comprehensive Assessment*, MIT Press.

Chalmers, David, 1995, "Facing up to the problem of consciousness," *Journal of Consciousness Studies*, 2, pp200-19.

1996, *The Conscious Mind*, Oxford Univ. Press.

1999, (1999, p8.) "First-person Methods in the Science of Consciousness," *Consciousness Bulletin*, Fall 1999, pp8-11.

Crick, Francis, 1994, *The Astonishing Hypothesis: the Scientific Search for the Soul*, New York: Scribner.

Damasio, Antonio, 1999, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, New York: Harcourt Brace.

Debner, J. A., and Jacoby, 1994, "Unconscious perception: Attention, awareness and control," *Journal of Experimental Psychology: Learning Memory and Cognition*, 20, pp304-317.

Dennett, Daniel, 1988, "Quining Qualia," in A. Marcel and E. Bisiach, eds, *Consciousness in Modern Science*, Oxford University Press, pp42-77.

1991, *Consciousness Explained*, Boston: Little, Brown, and London: Allen Lane.

1994a, "Real Consciousness" in *Consciousness in Philosophy and Cognitive Neuroscience*, A. Revonsuo and M. Kamppinen, eds., Hillsdale, NJ: Lawrence Erlbaum, 1994, pp55-63.

1994b, "Instead of Qualia" in *Consciousness in Philosophy and Cognitive Neuroscience*, A. Revonsuo and M. Kamppinen, eds., Hillsdale, NJ: Lawrence Erlbaum, 1994. pp129-139.

1995, "Overworking the Hippocampus," commentary on Jeffrey Gray, *Behavioral and Brain Sciences*, 18, no. 4, 1995, pp677-78.

1996a, "Consciousness: More like Fame than Television," (in German translation) "Bewusstsein hat mehr mit Ruhm als mit Fernsehen zu tun," Christa Maar, Ernst Pöppel, and Thomas Christaller, eds., *Die Technik auf dem Weg zur Seele*, Munich: Rowohlt, 1996.

1996b, "Facing Backwards on the Problem of Consciousness," commentary on Chalmers for *Journal of Consciousness Studies*, 3, no. 1 (special issue, part 2), 1996, pp4-6, reprinted in *Explaining Consciousness—The 'Hard Problem,'* ed. Jonathan Shear, MIT Press/A Bradford Book, 1997.

1998a, "The Myth of Double Transduction," in the volume of the *International Consciousness Conference*, *Toward a Science of Consciousness II, The Second Tucson Discussions and Debates*, S. Hameroff, ed., A.W. Kaszniak, and A.C. Scott, MIT Press, 1998, pp97-107.

1998b, "No bridge over the stream of consciousness," Commentary on Pessoa et al: in *Behavioral and Brain Sciences*, 21, pp753-54.

Forthcoming, "The Zombic Hunch: The Extinction of an Illusion?" in *Philosophy* (special issue on philosophy at the Millennium).

Dennett and Kinsbourne, M., 1992, "Time and the Observer: The Where and When of Consciousness in the Brain," *Behavioral and Brain Sciences*, 15, pp183-247, 1992.

Fodor, Jerry, 1998, "Review of Steven Pinker's *How the Mind Works*, and Henry Plotkin's *Evolution in Mind*," *London Review of Books*, Jan 22, 1998, reprinted in Fodor, *In Critical Condition*, 1998, Bradford Book/MIT Press.

Humphrey, Nicholas, 2000, "How to Solve the Mind-Body Problem" (with commentaries and a reply by the author), *Journal of Consciousness Studies,* 7, pp5-20.

Hurley, Susan, 1998, *Consciousness in Action*, Cambridge, MA: Harvard University Press.

McCarthy, John, 1959, discussion of Oliver Selfridge, "Pandemonium: A Paradigm for Learning," *Symposium on the Mechanization of Thought Processes*, London: HM Stationery Office.

Penrose, Roger, 1989, *The Emperors New Mind: Concerning Computers, Minds and the Laws of Physics*, Oxford: Oxford Univ. Press.

Pessoa, L., Thompson, E., and Noë, A., 1998, "Finding Out About Filling In: A Guide to Perceptual Completion for Visual Science and the Philosophy of Perception," *Behavioral and Brain Sciences*, 21, pp723-802.

Petitot, J., Varela, F, Pachoud, B., and Roy, J-M., 1999, *Naturalizing Phenomenology: Issues in Contemporary*

Phenomenology and Cognitive Science, Stanford, CA: Stanford Univ. Press.
Rozin, Paul, 1976, "The Evolution of Intelligence and Access to the Cognitive Unconscious," *Progress in Psychobiology and Physiological Psychology*, 6, New York: Academic Press, pp245-80.
Searle, John, 1980, "Minds, Brains, and Programs," *Behavioral and Brain Sciences,* 3, pp417-58.
Voorhees, Burton, 2000, "Dennett and the Deep Blue Sea," *Journal of Consciousness Studies,* 7, pp53-69.
Weiskrantz, L., 1998, "Consciousness and Commentaries," in Hameroff, SR, Kaszniak, A. W., and Scott, A. C., eds. *Towards a Science of Consciousness II: The Second Tucson Discussions and Debates*, Cambridge, MA: MIT Press, pp11-25.
Wright, Robert, 2000, *Nonzero: the Logic of Human Destiny,* New York: Pantheon.

Consciousness: How Much Is That in Real Money?

for R. Gregory, ed., *Oxford Companion to the Mind*, on consciousness. December 12, 2001

Consciousness often seems to be utterly mysterious. I suspect that the principal cause of this bafflement is a sort of accounting error that is engendered by a familiar series of challenges and responses. A simplified version of one such path to mysteryland runs as follows:

Phil: What is consciousness?

Sy: Well, some things–such as stones and can-openers—are utterly lacking in any *point of view*, any *subjectivity* at all, while other things—such as you and me—do have points of view: private, perspectival, interior ways of being apprised of some limited aspects of the wider world and our bodies' relations to it. We lead our lives, suffering and enjoying, deciding and choosing our actions, guided by this "first-person" access that we have. To be conscious is to be an agent with a point of view.

Phil: But surely there is more to it than that! A cherry tree has limited access to the ambient temperature at its surface, and can be (mis-)guided into blooming inopportunely by unseasonable warm weather; a robot with video camera "eyes" and microphone "ears" may discriminate and respond aptly to hundreds of different aspects of its wider world; my own immune system can sense, discriminate, and respond appropriately (for the most part) to millions of different eventualities. Each of these is an agent (of sorts) with a point of view (of sorts) but none of them is conscious.

Sy: Yes, indeed; there is more. We conscious beings have capabilities these simpler agents lack. We don't just notice things and respond to them; we *notice* that we notice things. More exactly, among the many discriminative states that our bodies may enter (including the states of our immune systems, our autonomic nervous systems, our digestive systems, and so forth), a subset of them can be discriminated in turn by higher-order discriminations which then become sources of guidance for higher level control activities. In us, this recursive capacity for self-monitoring exhibits no clear limits—beyond those of available time and energy. If somebody throws a brick at you, you see it coming and duck. But you also discriminate the fact that you *visually* discriminated the projectile, and can then discriminate the further fact that you can tell visual from tactile discriminations (usually), and then go on to reflect on the fact that you are also able to recall recent sensory discriminations in some detail, and that there is a difference between experiencing something and recalling the experience of something, and between thinking about the difference between recollection and experience and thinking about the difference between seeing and hearing, and so forth, till bedtime.

Phil: But surely there is more to it than that! Although existing robots may have quite paltry provisions for such recursive self-monitoring, I can readily imagine this particular capacity being added to some robot of the future. However deftly it exhibited its capacity to generate and react appropriately to "reflective" analyses of its underlying discriminative states, it wouldn't be conscious—not the way we are.

Sy: Are you sure you can imagine this?

Phil: Oh yes, absolutely sure. There would be, perhaps, some sort of *executive* point of view definable by analysis of the power such a robot would have to control itself based on these reactive capacities, but this robotic subjectivity would be a pale shadow of ours. When it uttered "it seems to me …" its utterances wouldn't really mean anything—or at least, they wouldn't mean what I mean when I tell you what it's like to be me, how things seem to me.

Sy: I don't know how you can be so confident of that, but in any case, you're right that there is more to consciousness than that. Our discriminative states are not just discriminable; they have the power to provoke preferences in us. Given choices between them, we are not

indifferent, but these preferences are themselves subtle, variable, and highly dependent on other conditions. There is a time for chocolate and a time for cheese, a time for blue and a time for yellow. In short (and oversimplifying hugely), many if not all of our discriminative states have what might be called a dimension of affective valence. We care which states we are in, and this caring is reflected in our dispositions to change state.

Phil: But surely there is more to it than that! When I contemplate the luscious warmth of the sunlight falling on that old brick wall, it's not just that I prefer looking at the bricks to looking down at the dirty sidewalk beneath them. I can readily imagine outfitting our imaginary robot with built-in preferences for every possible sequence of its internal states, but it would still not have anything like my conscious *appreciation* of the visual poetry of those craggy, rosy bricks.

Sy: Yes, I grant it; there is more. For one thing, you have meta-preferences; perhaps you wish you could stop those sexual associations from interfering with your more exalted appreciation of the warmth of that sunlight on the bricks, but at the same time (roughly) you are delighted by the persistence of those saucy intruders, distracting as they are, but . . . what was it you were trying to think about? Your stream of consciousness is replete with an apparently unending supply of associations. As each fleeting occupant of the position of greatest influence gives way to its successors, any attempt to halt this helter-skelter parade and monitor the details of the associations only generates a further flood of evanescent states, and so on. Coalitions of themes and projects may succeed in dominating "attention" for some useful and highly productive period of time, fending off would-be digressions for quite a while, and creating the sense of an abiding self or ego taking charge of the whole operation. And so on.

Phil: But surely there is more to it than that! And now I begin to see what is missing from your deliberately evasive list of additions. All these dispositions and meta-dispositions to enter into states and meta-states and meta-meta-states of reflection about reflection could be engineered (I dimly imagine) into some robot. The trajectory of its internal state-switching could, I suppose, look strikingly similar to the "first-person" account I might give of my own stream of consciousness, but those states of the robot would have no actual *feel*, no *phenomenal* properties at all! You're still leaving out what the philosophers call the *qualia*.

Sy: Actually, I'm still leaving out *lots* of properties. I've hardly begun acknowledging all the oversimplifications of my story so far, but now you seem to want to pre-empt any further additions from me by insisting that there are properties of consciousness that are altogether different from the properties I've described so far. I thought I *was* adding "phenomenal" properties in response to your challenge, but now you tell me I haven't even begun. Before I can tell if I'm leaving these properties out, I have to know what they are. Can you give me a clear example of a phenomenal property? For instance, if I used to like a particular shade of yellow, but thanks to some traumatic experience (I got struck by a car of that color, let's suppose) that shade of yellow now makes me very uneasy (whether or not it reminds me explicitly of the accident), would this suffice to change the *phenomenal* properties of my experience of that shade of yellow?

Phil: Not necessarily. The *dispositional* property of making you uneasy is not itself a phenomenal property. Phenomenal properties are, by definition, not dispositional but rather intrinsic and accessible only from the first-person point of view . . .

Thus we arrive in mysteryland. If you define *qualia* as *intrinsic properties* of experiences considered in isolation from all their causes and effects, logically independent of all dispositional properties, then they are logically guaranteed to elude all broad functional analysis—but it's an empty victory, since there is no reason to believe such properties exist. To see this, compare the *qualia* of experience to the *value* of money. Some naive Americans can't get it out of their heads that dollars, unlike francs and marks and yen, have *intrinsic value* ("How much is that in *real* money?"). They are quite content to "reduce" the value of other currencies in dispositional terms to their exchange rate with dollars (or goods and services), but they have a hunch that dollars are different. Every dollar, they declare, has something logically independent of its functionalistic exchange powers, which we might call

its *vim*. So defined, the *vim* of each dollar is guaranteed to elude the theories of economists forever, but we have no reason to believe in it—aside from the heartfelt hunches of those naive Americans, which can be explained without being honored.

Some participants in the consciousness debates simply demand, flat out, that their intuitions about phenomenal properties are a non-negotiable starting point for any science of consciousness. Such a conviction must be considered an interesting symptom, deserving a diagnosis, a datum that any science of consciousness must account for, in the same spirit that economists and psychologists might set out to explain why it is that so many people succumb to the potent illusion that money has intrinsic value.

There are many properties of conscious states that can and should be subjected to further scientific investigation right now, and once we get accounts of them in place, we may well find that they satisfy us as an explanation of what consciousness is. After all, this is what has happened in the case of the erstwhile mystery of what *life* is. Vitalism—the insistence that there is some big, mysterious extra ingredient in all living things—turns out to have been not a deep insight but a failure of imagination. Inspired by that happy success story, we can proceed with our scientific exploration of consciousness. If the day arrives when all these acknowledged debts are paid and we plainly see that something big is missing (it should stick out like a sore thumb at some point, if it is really important) those with the unshakable hunch will get to say they told us so. In the meantime, they can worry about how to fend off the diagnosis that they, like the vitalists before them, have been misled by an illusion.

Who's on First? Heterophenomenology Explained

There is a pattern of miscommunication bedeviling the people working on consciousness that is reminiscent of the classic Abbott and Costello 'Who's on First?' routine. With the best of intentions, people are talking past each other, seeing major disagreements when there are only terminological or tactical preferences—or even just matters of emphasis—that divide the sides. Since some substantive differences also lurk in this confusion, it is well worth trying to sort out. Much of the problem seems to have been caused by some misdirection in my apologia for *heterophenomenology* (Dennett, 1982; 1991), advertised as an explicitly *third*-person approach to human consciousness, so I will try to make amends by first removing those misleading signposts and sending us back to the real issues.

On the face of it, the study of human consciousness involves phenomena that seem to occupy something rather like another dimension: the private, subjective, '*first*-person' dimension. Everybody agrees that this is where we start. What, then, is the relation between the standard 'third-person' objective methodologies for studying meteors or magnets (or human metabolism or bone density), and the methodologies for studying human consciousness? Can the standard methods be extended in such a way as to do justice to the phenomena of human consciousness? Or do we have to find some quite radical or revolutionary alternative science? I have defended the hypothesis that there is a straightforward, conservative extension of objective science that handsomely covers the ground—*all* the ground—of human consciousness, doing justice to all the data without ever having to abandon the rules and constraints of the experimental method that have worked so well in the rest of science. This third-person methodology, dubbed heterophenomenology (phenomenology *of another,* not oneself), is, I have claimed, the sound way to take the *first*-person point of view as seriously as it can be taken.

To place heterophenomenology in context, consider the following ascending scale of methods of scientific investigation:

experiments conducted on anaesthetized animals;
experiments conducted on awake animals;
experiments on human subjects conducted in 'behaviorese'
—subjects are treated as much as possible like laboratory rats, trained to criterion with the use of small rewards, with minimal briefing and debriefing, etc.;
experiments in which human subjects collaborate with experimenters
—making suggestions, interacting verbally, telling what it is like.

Only the last of these methods holds out much hope of taking human subjectivity seriously, and at first blush it may seem to be a first-person (or, with its emphasis on communicative interaction with the subjects, second-person) methodology, but in fact it is *still* a third-person methodology if conducted properly. It is heterophenomenology.

Most of the method is so obvious and uncontroversial that some scientists are baffled that I would even call it a method: basically, you have to take the vocal sounds emanating from the subjects' mouths (and your own mouth) and *interpret* them! Well of course. What else could you do? Those sounds aren't just belches and moans; they're speech acts, reporting, questioning, correcting, requesting, and so forth. Using such standard speech acts, other events such as button-presses can be set up to be interpreted as speech acts as well, with highly specific meanings and fine temporal resolution. What this interpersonal communication enables you, the investigator, to do is to compose a catalogue of *what the subject believes to be true about his or her conscious experience*. This catalogue of beliefs fleshes out the subject's *heterophenomenological world*, the world according to S—the subjective world of one subject—not to be confused with the real world. The total set of details of heterophenomenology, plus all the data we can gather about concurrent events in the brains of subjects and in the surrounding environment, comprise the total data set for a theory of human consciousness.

It leaves out no objective phenomena and no subjective phenomena of consciousness.

Just what kinds of things does this methodology commit us to? Beyond the unproblematic things all of science is committed to (neurons and electrons, clocks and microscopes...) just to *beliefs*—the beliefs expressed by subjects and deemed constitutive of their subjectivity. And what kind of things are beliefs? Are they sentences in the head written in brain writing? Are they nonphysical states of dualist ectoplasm? Are they structures composed of proteins or neural assemblies or electrical fields? We may stay maximally noncommittal about this by adopting, at least for the time being (I recommend: for ever), the position I have defended (Dennett, 1971; 1987; 1991) that treats beliefs from *the intentional stance* as *theorists' fictions* similar to centres of mass, the equator, and parallelograms of forces. In short, we may treat beliefs as *abstractions* that measure or describe the complex cognitive state of a subject rather the way horsepower indirectly but accurately measures the power of engines (don't look in the engine for the horses). As Churchland (1979) has pointed out, physics already has hundreds of well-understood measure predicates, such as *x has weight-in-grams n*, or *x is moving up at n meters per second*, which describe a physical property of x by relating it to a *number*. Statements that attribute beliefs using the standard *propositional attitude* format, *x believes that p*, describe x's internal state by relating it to a *proposition*, another kind of useful abstraction, systematized in logic, not arithmetic. We need beliefs anyway for the rest of social science, which is almost entirely conducted in terms of the intentional stance, so this is a conservative exploitation of already quite well-behaved and well-understood methods.

A catalogue of beliefs about experience is not the same as a catalogue of experiences themselves, and it has been objected (Levine, 1994) that 'conscious experiences themselves, not merely our verbal judgments about them, are the primary data to which a theory must answer.' But how, in advance of theory, could we catalogue the experiences themselves? We can see the problem most clearly in terms of a nesting of proximal sources that are presupposed as we work our way up from raw data to heterophenomenological worlds:

(a) 'conscious experiences themselves'
(b) beliefs about these experiences
(c) 'verbal judgments' expressing those beliefs
(d) utterances of one sort or another

What are the 'primary data'? For heterophenomenologists, the *primary* data are the utterances, the *raw*, uninterpreted data. But before we get to theory, we can interpret these data, carrying us via (c) speech acts to (b) beliefs about experiences.[1] These are the primary *interpreted* data, the pretheoretical data, the *quod erat explicatum* (as organized into heterophenomenological worlds), for a science of consciousness. In the quest for primary data, Levine wants to go all the way to (a) conscious experiences themselves, instead of stopping with (b) subjects' beliefs about their experiences, but this is not a good idea. If (a) outruns (b)—if you have conscious experiences you don't believe you have—those extra conscious experiences are just as inaccessible *to you* as to the external observers. So Levine's proposed alternative garners you no more usable data than heterophenomenology does. Moreover, if (b) outruns (a)—if you believe you have conscious experiences that you don't in fact have—then it is your beliefs that we need to explain, not the non-existent experiences! Sticking to the heterophenomenological standard, then, and treating (b) as the maximal set of primary data, is the way to avoid any commitment to spurious data.

But what if some of your beliefs are inexpressible in verbal judgments? If you believe *that*, you can tell us, and we can add that belief to the list of beliefs in our primary data: 'S claims that he has ineffable beliefs about X.' If this belief is true, then we encounter the obligation to explain what these beliefs are and why they are ineffable. If this belief is false, we still have to explain why S believes (falsely) that there are these particular ineffable beliefs. As I put it in *Consciousness Explained*,

> You are *not* authoritative about what is happening in you, but only about what *seems* to be happening in you, and we are giving you total, dictatorial authority over the account of how it seems to you, about *what it is like to be you*. And if you complain that some parts of how it seems to you are ineffa-

[1] Doesn't interpretation require theory? Only in the minimal sense of presupposing that the entity interpreted is an intentional system, capable of meaningful communication. The task of unifying the interpretation of all the verbal judgments into a heterophenomenological world is akin to reading a novel, in contrast to reading what purports to be true history or biography. The issue of truth and evidence does not arise, and hence the interpretation is as neutral as possible between different theories of what is actually happening in the subject.

> ble, we heterophenomenologists will grant that too. What better grounds could we have for believing that you are unable to describe something than that (1) you don't describe it, and (2) confess that you cannot? Of course you might be lying, but we'll give you the benefit of the doubt (Dennett, 1991, pp. 96—7).

This is all quite obvious, but it has some under-appreciated implications. Exploiting linguistic communication in this way, you get a fine window into the subject's subjectivity but at the cost of a peculiar lapse in normal interpersonal relations. You *reserve judgment* about whether the subject's beliefs, as expressed in their communication, are true, or even well-grounded, but then you treat them as *constitutive* of that subject's subjectivity. (As far as I can see, this is the third-person parallel to Husserl's notion of bracketing or *epoché*, in which the normal presuppositions and inferences of one's own subjective experience are put on hold, as best one can manage, in order to get at the core experience, as theory-neutral and unencumbered as possible.) This interpersonal reserve can be somewhat creepy. To put it fancifully, suppose you burst into my heterophenomenology lab to warn me that the building is on fire. I don't leap to my feet and head for the door; I write down 'subject S believes the building is on fire.' 'No, really, it's on fire!' you insist, and I ask, 'Would you like to expand on that? *What is it like* for you to think the building is on fire?' and so forth. In one way I am taking you as seriously as you could ever hope to be taken, but in another way I am not. I am not *assuming* that you are right in what you tell me, but just that that is what you do believe. Of course most of the data-gathering is not done by any such simple interview. Experiments are run in which subjects are prepared by various conversations, hooked up to all manner of apparatus, etc., and carefully debriefed. In short, heterophenomenology is nothing new; it is nothing other than the method that has been used by psychophysicists, cognitive psychologists, clinical neuropsychologists, and just about everybody who has ever purported to study human consciousness in a serious, scientific way.

This point has sometimes been misunderstood by scientists who suppose, quite reasonably, that since I am a philosopher I must want to scold somebody for something, and hence must be proposing restrictions on standard scientific method, or discovering limitations therein. On the contrary, I am urging that the prevailing methodology of scientific investigation on human consciousness is not only sound, but readily extendable in non-revolutionary ways to incorporate *all* the purported exotica and hard cases of human subjectivity. I want to put the burden of proof on those who insist that third-person science is incapable of grasping the nettle of consciousness.

Let me try to secure the boundaries of the heterophenomenological method more clearly, then, since this has apparently been a cause of confusion. As Anthony Jack has said to me:

> It strikes me that heterophenomenology is a method in the same way that 'empiricism' is a method, but no more specific nor clearly defined than that. Given how general you seem to allow your definition of heterophenomenology to be, it is no surprise that everything conforms! Perhaps it would be clearer if you explained more clearly what it is supposed to be a counterpoint to—what it is that you object to. I know I am not the only one who has a feeling that you make the goalposts surprisingly wide. So what exactly is a foul? (Jack, personal correspondence).

Lone-wolf autophenomenology, in which the subject and experimenter are one and the same person, is a foul, not because you can't do it, but because it isn't science until you turn your self-administered pilot studies into heterophenomenological experiments. It has always been good practice for scientists to put themselves in their own experimental apparatus as informal subjects, to confirm their hunches about what it feels like, and to check for any overlooked or underestimated features of the circumstances that could interfere with their interpretations of their experiments. But scientists have always recognized the need to confirm the insights they have gained from introspection by conducting properly controlled experiments with naive subjects. As long as this obligation is met, whatever insights one may garner from 'first-person' investigations fall happily into place in 'third-person' heterophenomenology. Purported discoveries that cannot meet this obligation may inspire, guide, motivate, illuminate one's scientific theory, but *they* are not data—the beliefs of subjects about them are the data. Thus if some phenomenologist becomes convinced by her own (first-)personal experience, however encountered, transformed, reflected upon, of the existence of a feature of consciousness in need of explanation and accommo-

dation within her theory, her *conviction that this is so* is itself a fine datum in need of explanation, by her or by others, but the truth of her conviction must not be presupposed by science.

Does anybody working on consciousness disagree with this? Does anybody think that one can take personal introspection by the investigator as constituting stand-alone evidence (publishable in a peer-reviewed journal, etc.) for any substantive scientific claim about consciousness? I don't think so. It is taken for granted, so far as I can see, by all the authors in this volume that there is no defensible 'first-person science' lying in this quarter, even though that would be the most obvious meaning of the phrase 'taking a first-person approach.' Thus Cytowic, and Hubbard and Jack, discuss the difficulties in confirming that synaesthesia is more or less what synaesthetes say it is, and never question the requirement that 'taking the phenomenological reports of these subjects seriously' (Hubbard and Jack, abstract) requires 'the personal interaction between subject and experimenter.' And when Hurlburt and Heavey say (abstract), 'For example, first-person investigators often rely on questions such as "What were you thinking when you...?" or "How were you feeling when you…?"' it apparently does not occur to them that these *aren't* first-person investigations; they are third-person investigations of the special kind that exploit the subject's capacity for verbal communication. They are heterophenomenological inquiries. So I think we can set aside lone-wolf autophenomenology in all its guises. It is not an attractive option, for familiar reasons. The experimenter/subject duality is not what is being challenged by those who want to go beyond the 'third-person' methodology. What other alternatives should we consider?

Several critics have supposed that heterophenomenology, as I have described it, is too agnostic or too neutral. Goldman (1997) says that heterophenomenology is not, as I claim, the standard method of consciousness research, since researchers 'rely substantially on subjects' introspective beliefs about their conscious experience (or lack thereof)' (p. 532). In personal correspondence (Feb 21, 2001, available as part of my debate with Chalmers, on my website, at http://ase.tufts.edu/cogstud/papers/chalmersdeb3dft.htm) he puts the point this way:

> The objection lodged in my paper [Goldman, 1997] to heterophenomenology is that what cognitive scientists *actually* do in this territory is not to practice agnosticism. Instead, they rely substantially on subjects' introspective beliefs (or reports). So my claim is that the heterophenomenological method is not an accurate description of what cognitive scientists (of consciousness) standardly do. Of course, you can say (and perhaps intended to say, but if so it wasn't entirely clear) that this is what scientists *should* do, not what they *do* do.

I certainly would play the role of reformer if it were necessary, but Goldman is simply mistaken; the adoption of agnosticism is so firmly built into practice these days that it goes without saying, which is perhaps why he missed it. Consider, for instance, the decades-long controversy about mental imagery, starring Roger Shepard, Steven Kosslyn, and Zenon Pylyshyn among many others. It was initiated by the brilliant experiments by Shepard and his students in which subjects were shown pairs of line drawings like the pair in figure 1, and asked to press one button if the figures were different views of the same object (rotated in space) and another button if they were of different objects. Most subjects claim to solve the problem by rotating one of the two figures in their 'mind's eye' or imagination, to see if it could be superimposed on the other. Were subjects really doing this 'mental rotation'? By varying the angular distance

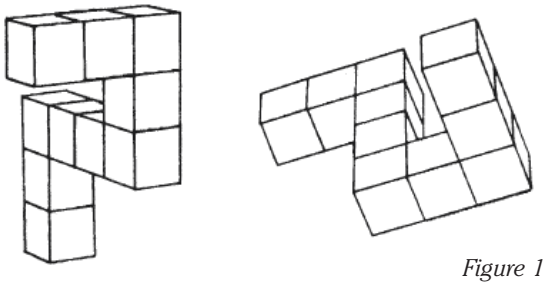


Figure 1

actually required to rotate the two figures into congruence, and timing the responses, Shepard was able to establish a remarkably regular linear relation between latency of response and angular displacement. Practiced subjects, he reported, are able to rotate such mental images at an angular velocity of roughly 60E per second (Shepard and Metzler, 1971). This didn't settle the issue, since Pylyshyn and others were quick to compose alternative hypotheses that could account for this striking temporal relationship. Further studies were called for and executed, and the controversy continues to generate new experiments and analysis today (see Pylyshyn, forthcoming, for an excellent survey of the

history of this debate; also my commentary, Dennett, forthcoming, both in *Behavioral and Brain Sciences*). Subjects always *say* that they are rotating their mental images, so if agnosticism were not the tacit order of the day, Shepard and Kosslyn would have never needed to do their experiments to support subjects' claims that what they were doing (at least if described metaphorically) really was a process of image manipulation. Agnosticism is built into all good psychological research with human subjects. In psychophysics, for instance, the use of signal detection theory has been part of the canon since the 1960s, and it specifically commands researchers to control for the fact that the response criterion is under the subject's control although the subject is not himself or herself a reliable source on the topic. Or consider the voluminous research literature on illusions, both perceptual and cognitive, which standardly assumes that the data are what subjects judge to be the case, and never makes the mistake of 'relying substantially on subjects' introspective beliefs.'

The diagnosis of Goldman's error is particularly clear here: of course experimenters on illusions rely on subjects' introspective beliefs (as expressed in their judgments) about how it *seems* to them, but that *is* the agnosticism of heterophenomenology; to go beyond it would be, for instance, to assume that in size illusions there really were visual images of different sizes somewhere in subjects' brains (or minds), which of course no researcher would dream of doing.[2]

David Chalmers has recently made a similar, if vaguer, claim:

> Dennett... says scientists have to take a neutral attitude (taking reports themselves as data, but making no claims about their truth), because reports can go wrong. But this misses the natural intermediate option that Max Velmans has called critical phenomenology: accept verbal reports as a prima facie guide to a subject's conscious experience, except where there are specific reasons to doubt their reliability. This seems to be most scientists' attitude toward verbal reports and consciousness: it's not 'uncritical acceptance,' but it's also far from the 'neutrality' of heterophenomenology (Chalmers, 2003).

Chalmers neglects to say how Velmans' critical phenomenology is 'far from' the neutrality of heterophenomenology. I conducted a lengthy correspondence with Velmans on this score and was unable to discover what the purported difference is, beyond Velmans' insisting that his method 'accepts the reality of first-person experience,' but since it is unclear what this means, this is something a good scientific method should be agnostic about. Neither Chalmers nor Velmans has responded to my challenge to describe an experiment that is licensed by, or motivated by, or approved by 'critical phenomenology' but off-limits to heterophenomenology, so if there is a difference here, it is one of style or emphasis, not substance. Chalmers has acknowledged this, in a way:

> Dennett 'challenges' me to name an experiment that 'transcends' the heterophenomenological method. But of course both views can accommodate experiments equally: every time I say we're using a verbal report or introspective judgment as a guide to first-person data, he can say we're using it as third-person data, and vice versa. So the difference between the views doesn't lie in the range of experiments 'compatible' with them. Rather, it lies in the way that experimental results are interpreted. And I think the interpretation I'm giving (on which reports are given prima facie credence as a guide to conscious experience) is by far the most common attitude among scientists in the field. Witness the debate about unconscious perception among cognitive psychologists about precisely which third-person measures (direct report, discrimination, etc.) are the best guide to the presence of conscious perception. Here, third-person data are being used as a (fallible) guide to first-person data about consciousness, which are of primary interest. On the heterophenomenological view, this debate is without much content: some states subserve report, some subserve discrimination, etc., and that's about all there is to say. I think something like this is Dennett's attitude to those debates, but it's not the attitude of most of the scientists working in the field (Chalmers, 2003).

Chalmers misconstrues my view, as we can see if we look more closely at a particular debate about uncon-

[2] Goldman has responded to this paragraph in a series of emails to me, which I have included in an Appendix on the website mentioned above.

scious perception, to see how heterophenomenology sorts out the issues. Consider *masked priming*. It has been demonstrated in hundreds of different experiments that if you present subjects with a 'priming' stimulus, such as a word or picture flashed briefly on a screen in front of the subject, followed very swiftly by a 'mask'—a blank or sometimes randomly patterned rectangle—before presenting the subjects with a 'target' stimulus to identify or otherwise respond to, there are conditions under which subjects will manifest behaviour that shows they have disciminated the priming stimulus, while they candidly and sincerely report that they were entirely unaware of any such stimulus. For instance, asked to complete the word stem *fri___*, subjects who have been shown the priming stimulus *cold* are more likely to comply with *frigid* and subjects who have been shown the priming stimulus *scared* are more likely to comply with *fright* or *frightened*, even though both groups of subjects claim not to have seen anything but first a blank rectangle followed by the target to be completed. Now are subjects to be trusted when they say that they were not conscious of the priming stimulus? There are apparently two ways theory can go here:

A. Subjects are conscious of the priming stimulus and then the mask makes them immediately forget this conscious experience, but it nevertheless influences their later performance on the target.
B. Subjects unconsciously extract information from the priming stimulus, which is prevented from 'reaching consciousness' by the mask.

Chalmers suggests that it is my 'attitude' that there is nothing to choose between these two hypotheses, but my point is different. It is open for scientific investigation to develop reasons for preferring one of these theoretical paths to the other, but *at the outset*, heterophenomenology is neutral, leaving the subject's heterophenomenological worlds bereft of any priming stimuli—that is how it seems to the subjects, after all—while postponing an answer to the question of how or why it seems thus to the subjects. Heterophenomenology is the beginning of a science of consciousness, not the end. It is the organization of the data, a catalogue of *what must be explained*, not itself an explanation or a theory. (This was the original meaning of 'phenomenology': a pretheoretical catalogue of the phenomena the theory must account for.) And in maintaining this neutrality, it is actually doing justice to the *first-person* perspective, because you yourself, as a subject in a masked priming experiment, cannot discover anything in your experience that favours A or B. (If you think you can discover something—if you notice some glimmer of a hint in the experience, speak up! You're the subject, and you're supposed to tell it like it is. Don't mislead the experimenters by concealing something you discover in your experience. Maybe they've set the timing wrong for you. Let them know. But if they've done the experiment right, and you really find, so far as you can tell from your own first-person perspective, that you were not conscious of any priming stimulus, then say so, and note that both A and B are still options between which you are powerless to offer any further evidence.)

But now suppose scientists look for a good reason to favour A or B and find it. What could it be? A theory that could provide a good reason would be one that is well-confirmed in other domains or contexts and that distinguishes, say, the *sorts* of discriminations that can be made unconsciously from the sorts that require consciousness. If in this case the masked discrimination was of a feature that in all other circumstances could only be discriminated by a conscious subject, this would be a (fairly) good reason for supposing that, however it may be with other discriminations, in this case the discrimination was conscious-and-then-forgotten, not unconscious. Notice that if anything at all like this were discovered, and used as a ground for distinguishing A from B, it would be a triumph of *third-person* science, not due to anything that is accessible only to the subject's introspection. Subjects would *learn for the first time* that they were, or were not, conscious of these stimuli when they were taught the theory. It is the neutrality of heterophenomenology that permits such a question to be left open, pending further development of theory. And of course anyone proposing such a theory would have to have bootstrapped their way to their own proprietary understanding of what they meant by conscious and unconscious subjects, finding a consilience between our everyday assumptions about what we are conscious of and what we are not, on the one hand, and their own classificatory scheme on the other. Anything too extreme ('It turns out on our theory that most people are conscious for only a few seconds a day, and nobody is conscious of sounds at all; hearing is entirely unconscious perception') will be rightly dismissed as an abuse of common understanding of the terms, but a theory that is predictively fecund and ele-

gant can motivate substantial abandonment of this anchoring lore. Only when such a theory is in place will we be able, for the first time, to *know what we mean* when we talk about 'the experiences themselves' as distinct from what we each, subjectively, take our experiences to be.

This sketches a clear path to settling the issue between A and B, or to discovering good reasons for declaring the question ill-posed. If Chalmers thinks that scientists do, and should, prefer a different attitude towards such questions, he should describe in some detail what it is and why it is preferable. In fact, I think that while there has been some confusion on this score (and some spinning of wheels about just what would count as favouring unconscious perception over conscious perception with forgetting), scientists are comfortable with the heterophenomenological standards.

Varela and Shear (1999) describe the empathy of the experimenter that they see as the distinguishing feature of a method they describe as first-person:

> In fact, that is how he sees his role: as an empathic resonator with experiences that are familiar to him and which find in himself a resonant chord. This empathic position is still partly heterophenomenological, since a modicum of critical distance and of critical evaluation is necessary, but the intention is entirely other: to meet on the same ground, as members of the same kind.... Such encounters would not be possible without the mediator being steeped in the domain of experiences under examination, as nothing can replace that first-hand knowledge. This, then, is a radically different style of validation from the others we have discussed so far (p. 10).

One can hardly quarrel with the recommendation that the experimenter be 'steeped in the domain of experiences' under examination, but is there more to this empathy than just good, knowledgeable interpretation? If so, what is it? In a supporting paper, Thompson speaks of 'sensual empathy,' and opines: 'Clearly, for this kind of sensual empathy to be possible, one's own body and the Other's body must be of a similar type' (2001, p. 33). This may be clear to Thompson, but in fact it raises a highly contentious set of questions: Can women not conduct research on the consciousness of men? Can slender investigators not explore the phenomenology of the obese? Perhaps more to the point, can researchers with no musical training or experience ('tin ears') effectively conduct experiments on the phenomenology of musicians? When guidance from experts is available, one should certainly avail oneself of it, but the claim that one must *be* an expert (an expert musician, an expert woman, an expert obese person) before conducting the research is an extravagant one. Suppose, however, that it is true. If so, we should be able to discover this by attempting, and detectably failing, to conduct the research as well as the relevant experts conduct the research. *That* discovery would itself be something that could only be made by first adopting the neutral heterophenomenological method and then assaying the results in comparison studies. So once again, the neutral course to pursue is not to *assume* that men can't investigate the consciousness of women, etc., but to investigate the question of whether we can discover any good scientific reason to believe this. If we can, then we should adjust the standards of heterophenomenology accordingly. It is just common sense to design one's experiments in such a way as to minimize interference and maximize efficiency and acuity of data-gathering.

Is there, then, any 'radically different style of validation' on offer in these proposals? I cannot find any. Some are uneasy about the noncommital stance of the heterophenomenologist. Wouldn't the cultivation of deep trust between subject and experimenter be better? Apparently not. The history of *folie à deux* and Clever Hans phenomena suggests that quite unwittingly the experimenter and the subject may reinforce each other into artifactual mutual beliefs that evaporate when properly probed. But we can explore the question. It is certainly wise for the experimenter not to antagonize subjects, and to encourage an atmosphere of 'trust'—note the scare quotes. The question is whether experimenters should go beyond this and *actually trust* their subjects, or should instead (as in standard experimental practice) quietly erect the usual barriers and foils that keep subjects from too intimate an appreciation of what the experimenters have in mind. Trust is a two-way street, surely, and the experimenter who gets in a position where the subject can do the manipulating has lost control of the investigation.

I suspect that some of the dissatisfaction with heterophenomenology that has been expressed is due to my not having elaborated fully enough the potential resources of this methodology. There are surely many subtleties of heterophenomenological method that

have yet to be fully canvassed. The policy of training subjects, in spite of its uneven history in the early days of psychology, may yet yield some valuable wrinkles. For instance, it might in some circumstances heighten the powers of subjects to articulate or otherwise manifest their subjectivity to investigators. The use of closed-loop procedures, in which subjects to some degree control the timing and other properties of the stimuli they receive, is another promising avenue. But these are not alternatives to heterophenomenology, which is, after all, just the conservative extension of standard scientific methods to data gathering from awake, communicating subjects.

Why *not* live by the heterophenomenological rules? It is important to appreciate that the reluctance to acquiesce in heterophenomenology as one's method is ideology-driven, not data-driven. Nobody has yet pointed to any variety of data that are inaccessible to heterophenomenology. Instead, they have objected 'in principle,' perhaps playing a little gorgeous Bach for the audience and then asking the rhetorical question, 'Can anybody seriously believe that the wonders of human consciousness can be exhaustively plumbed by *third-person methods??*' Those who are tempted to pose this question should either temper their incredulity for the time being or put their money where their mouth is by providing the scientific world with some phenomena that defy such methods, or by describing some experiments that are clearly worth doing but that would be ruled out by heterophenomenology. I suspect that some of the antagonism to heterophenomenology is generated by the fact that the very neutrality of the methodology opens the door to a wide spectrum of theories, including some—such as my own—that are surprisingly austere, deflationary theories according to which consciousness is more like stage magic than black magic, requiring no revolution in either physics or metaphysics. Some opponents to heterophenomenology seem intent on building the mystery into the very setting of the problem, so that such deflationary theories are disqualified at the outset. Winning by philosophical footwork what ought to be won by empirical demonstration has, as Bertrand Russell famously remarked, all the advantages of theft over honest toil. A more constructive approach recognizes the neutrality of heterophenomenology and accepts the challenge of demonstrating, empirically, in its terms, that there are marvels of consciousness that cannot be captured by conservative theories.

References

Chalmers, David J. (2003), 'Responses to articles on my work' http://www.u.arizona.edu/~chalmers/responses.html#dennett2).

Churchland, Paul M. (1979), *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press).

Dennett, Daniel C. (1971), 'Intentional Systems,' *J. Phil.*, 68, pp. 87–106.

Dennett, Daniel C. (1982), 'How to study consciousness empirically, or Nothing comes to mind,' *Synthese*, 59, pp. 159–80.

Dennett, Daniel C. (1987), *The Intentional Stance* (Cambridge, MA: MIT Press/Bradford).

Dennett, Daniel C. (1991), *Consciousness Explained* (Boston, MA: Little Brown).

Dennett, Daniel C. (forthcoming), 'Does your brain use the images in it, and if so, how?,' commentary on Pylyshyn (forthcoming).

Goldman, Alvin (1997), 'Science, Publicity and Consciousness,' *Philosophy of Science*, 64, pp. 525–45.

Levine, Joseph (1994), 'Out of the closet: A qualophile confronts qualophobia,' *Philosophical Topics*, 22, pp. 107–26.

Pylyshyn, Zenon W. (forthcoming), 'Mental Imagery: In search of a theory,' Target article in *Behavioral and Brain Sciences*.

Shepard, R.N., and Metzler, J. (1971), 'Mental rotation of three-dimensional objects,' *Science*, 171, pp. 701–3.

Thompson, Evan (2001), 'Empathy and consciousness,' *Journal of Consciousness Studies*, 8 (5–7), pp. 1–33.

Varela, Francisco, and Shear, Jonathan (1999), 'First-person methodologies: What, Why, How?,' *Journal of Consciousness Studies*, 6 (2–3), pp. 1–14.