14

Consciousness as an Evolved User-Illusion[98]

Keeping an open mind about minds

At last we are ready to put the pieces together and examine human consciousness as a system of virtual machines that evolved, genetically and memetically, to play very specialized roles in the "cognitive niche" our ancestors have constructed over the millennia. We are ready to confront Cartesian gravity head on and tackle some big questions:

1. How do human brains achieve "global" comprehension using "local" competences without invoking an intelligent designer?
2. Do our minds differ from the minds of other animals, and if so, how and why?

98 The Danish science journalist Tor Nørretranders published the Danish edition of his book *The User Illusion: Cutting Consciousness Down to Size*, in 1991, the same year my *Consciousness Explained* appeared, with my account of consciousness as a user-illusion. Neither of us was in a position to cite the other, obviously. The English translation of Nørretranders's book appeared in 1999. As I noted in my book (p. 311, fn. 9) among those who suggested somewhat similar forerunners of the idea were Kosslyn (1980), Minsky (1985), and Edelman (1989).

3. How did our manifest image become manifest to us?
4. Why do we experience things the way we do?

A brief review: Evolution has endowed all living things with the wherewithal to respond appropriately to their particular affordances, detecting and shunning the bad, detecting and obtaining the good, using the locally useful and ignoring everything else. This yields competence without comprehension, at every level from the molecular on up. Since there *can be* competence without comprehension, and since comprehension ("real" comprehension) is expensive, Nature makes heavy use of the Need to Know principle, and designs highly successful, adept, even cunning creatures who have no idea what they are doing or why. Reasons abound, but they are mostly free-floating rationales, undreamt of by those who benefit from them. As reverse engineers we can work out the ontology of affordances in the *Umwelt* of trees, fleas, and grizzly bears while remaining entirely agnostic about whether "it is like anything" to be them.

There can be reasons why trees do things without their *having* those reasons ("in mind"). Is there anything a flea can do for reasons that demonstrate that, unlike the tree, it *has* the reasons, it somehow "appreciates" the reasons that govern its actions? It might be that it is not "like anything" to be a flea, any more than it is like something to be an automatic elevator. And while we're at it, what makes us so sure it is like something to be a grizzly bear? It certainly *seems* to us that there must be something it is like to be a grizzly bear—just watch them, and listen to them! It seems more obvious that it is like something to be a grizzly bear than like the flea hiding in the grizzly bear's fur—doesn't it? But perhaps our imagination is playing tricks on us here. We know that it is like something to be *us* for the simple reason that we talk about it every day, in avowals, complaints, descriptions, poems, novels, philosophy books, and even peer-reviewed scientific papers. This is a central feature of our manifest image, *and that objective fact would be evident to any "Martian" scientists who studied us long enough to learn our languages.* Our introspective divulgences are behaviors that are just as observable and measurable as our acts of eating, running, fighting,

and loving. Is there anything we can do—aside from talking about it—that some other animals can also do that would clinch the case for their having consciousness more or less like ours? If the Martian scientists framed the question of whether other Earthlings—dolphins, chimpanzees, dogs, parrots, fish—were like the talking Earthlings, what would they point to, what would impress them, and why? That is not just a legitimate scientific question; it is an obligatory question, but one that thinkers commonly excuse themselves from addressing. They beg off, saying something along the lines of:

> I have no idea where, on the complexity scale of life, to draw the line—are worms, fish, reptiles, birds conscious? We may never know, but still we do know that we human beings are not alone in being conscious. That is obvious.

That is not acceptable, for two reasons. First, the idea that there is and must be a line drawn even if we don't know where to draw it is profoundly pre-Darwinian; there might be all manner of variations from poet to possum to peacock to perch to protozoon, and no "essence" of consciousness to discover. The fact that Nagel's (1974) famous formulation "what is it like?" is now treated as a crutch, a nudging gesture that lacks content but is presumed to point to the sought for (and cosmic) distinction, should be seen as an embarrassment, not an acceptable bit of temporizing—or even, by some philosophers, as a fundamental element of theory. Second, it ties our hands by appeal to an intuition—it is nothing more—that might be mistaken in some way. Temporary agnosticism about consciousness is fine—I have just advocated it—but not agnosticism saddled with the proviso that *of course other animals are conscious even if we can't say what that means.* That is, at best, the expression of confidence in the eventual triumph of the manifest image over the scientific image, in the face of a long history of defeats. It used to be obvious that the sun circled the earth, after all. People who won't let themselves even *think* about whether grizzly bears are "conscious like us" (whatever that means) are succumbing to ideology, not common sense. Their

motives may well be honorable—they are eager to expand the circle of the beings owed moral consideration because they can *suffer*—but until we can identify the important features, the features that matter, and explain why, this gesture is worse than empty. Worse, because it indefinitely postpones addressing the hard and important questions about exactly what suffering is, for instance, and whether insects or fish or daisies can suffer. We do have to "draw the line," for moral reasons, and most people (but not Jains, for instance) are comfortable with policies that call for peremptorily killing mosquitoes and deer ticks and tsetse flies, and *Plasmodium falciparum* (the microorganism that causes malaria—can a protozoan suffer?). Most people agree that rats can be exterminated but not their rodent cousins squirrels, who have been called rats with good PR by an insightful comedian. But we should keep our science neutral, just in case it surprises us with some important exceptions to folk wisdom.

And now, do you feel the pinch of Cartesian gravity? "Not like anything to be a grizzly bear? Are you kidding?" No, I am not kidding; I am putting the burden of demonstration on those who want to argue that some special phase shift occurs somewhere along the (apparent) continuum of growing adroitness that would put trees and fleas (or fleas and grizzly bears—take your choice) on different sides of a cosmic divide. There may be such a boundary, but unless "it's being like something" permits the organisms on one side to *do something important* (it might well be to *suffer*—but we'd have to find some objective way of detecting that) that they couldn't do if they were on the other side, it will be a boundary with nothing but folk tradition behind it. I am not denying the existence of such a boundary; I am postponing the issue, exploring how far we can get without postulating such a boundary, which is the way any scientific investigation should proceed. If you find yourself unable to tolerate that even-handedness, you are overcompensating for the effects of Cartesian gravity and disabling yourself from participation in the investigation. (It is worth remembering that Descartes solved the problem by fiat: only humans were conscious; animals were mindless automata.) If we shrink from his verdict, we still have to draw the *moral* line somewhere in the meantime, and let's err on the safe side, but unless we suspend *scientific* judgment until we have a better idea of what we're judging, we have no grounds for confirming, or moving, that line when we learn more. In the United Kingdom, the law since 1986 rules that the octopus (but only *Octopus vulgaris*—not any other cephalopods) is an "honorary vertebrate" entitled to legal protection. You may legally throw a live lobster or worm or moth into boiling water, for instance, but not an octopus; it has the same protection enjoyed by mammals and birds and reptiles. Should that law be expanded or retracted, or did the legislators get it right the first time? If we want a defensible answer to that question, we need to identify—and then bracket—our gut intuitions. We mustn't let our moral intuitions distort our empirical investigation from the outset.

The feats of clueless agents should not be underestimated. The termites' castle, the cuckoo chick's ovicide, and many other behavioral marvels are accomplished with only the sort of behavioral comprehension that amounts to practical know-how, unarticulated and unconsidered. When we human observers/explainers/predictors confront this well-designed excellence, we automatically set about figuring out the reasons why plants and animals do what they do, reverse engineering them with the aid of the intentional stance. And, as we have seen, when we do this it is common and natural to impute more understanding to an organism than it actually has, on the reasonable grounds that the behavior is manifestly clever, and whose cleverness is it, if not the organism's? Ironically, if we were creationists, we could comfortably attribute all the understanding to God and wouldn't feel so compelled to endow the organisms with it. They could all be God's marionettes. It was Darwin's discovery and exposure of the mindless process of natural selection, with its power to generate free-floating rationales, that freed our imaginations to continue reverse engineering all Nature's marvels without feeling an obligation to identify a mind that harbors the reasons we uncover.

How do human brains achieve "global" comprehension using "local" competences?

> Language was given to men so that they could conceal their thoughts.
>
> —Charles-Maurice de Talleyrand

> Language, like consciousness, only arises from the need, the necessity, of intercourse with others.
>
> —Karl Marx

> Consciousness generally has only been developed under the pressure of the necessity for communication.
>
> —Friedrich Nietzsche

There is no General Leslie Groves to organize and command the termites in a termite colony, and there is no General Leslie Groves to organize and command the even more clueless neurons in a human brain. How can human comprehension be composed of the activities of uncomprehending neurons? In addition to all the free-floating rationales that explain our many structures, habits, and other features, there are the anchored reasons we represent to ourselves and others. These reasons are themselves *things* for us, denizens of our manifest image alongside the trees and clouds and doors and cups and voices and words and promises that make up *our* ontology. We can *do things* with these reasons—challenge, reframe, abandon, endorse, disavow them—and these often covert behaviors would not be in our repertoires if we hadn't downloaded all the apps of language into our necktops. In short, we can think about these reasons, good and bad, and this permits them to influence our overt behaviors in ways unknown in other organisms.

The piping plover's distraction display or broken-wing dance gives the fox a reason to alter its course and approach her, but not by getting it to trust her. She may modulate her thrashing to hold the fox's attention, but the control of this modulation does not require her to have more than a rudimentary "appreciation" of the

fox's mental state. The fox, meanwhile, need have no more comprehension of just why it embarks on its quest instead of continuing to reconnoiter the area. We, likewise, can perform many quite adroit and *retrospectively* justifiable actions with only a vague conception of what we are up to, a conception often swiftly sharpened in hindsight by the self-attribution of reasons. It's this last step that is ours alone.

Our habits of self-justification (self-appreciation, self-exoneration, self-consolation, self-glorification, etc.) are ways of behaving (ways of *thinking*) that we acquire in the course of filling our heads with culture-borne memes, including, importantly, the habits of self-reproach and self-criticism. Thus we learn to plan ahead, to use the practice of reason-venturing and reason-criticizing to presolve some of life's problems, by talking them over with others and with ourselves. And not just talking them over—imagining them, trying out variations in our minds, and looking for flaws. We are not just Popperian but Gregorian creatures (see chapter 5), using thinking tools to design our own future acts. No other animal does that.

Our ability to do this kind of thinking is not accomplished by any dedicated brain structure not found in other animals. There is no "explainer-nucleus" for instance. Our thinking is enabled by the installation of a virtual machine made of virtual machines made of virtual machines. The goal of delineating and explaining this stack of competences via bottom-up neuroscience alone (without the help of *cognitive* neuroscience) is as remote as the goal of delineating and explaining the collection of apps on your smartphone by a bottom-up deciphering of its hardware circuit design and the bit-strings in memory without taking a peak at the user interface. The user interface of an app exists in order to make the competence accessible to users—people—who can't know, and don't need to know, the intricate details of how it works. The user-illusions of all the apps stored in our brains exist *for the same reason*: they make our competences (somewhat) accessible to users—*other* people—who can't know, and don't need to know, the intricate details. And then we get to use them ourselves, under roughly the same conditions, as guests in our own brains.

There might be some other evolutionary path—genetic, not cultural—to a somewhat similar user-illusion in other animals, but I have not been able to conceive of one in convincing detail, and according to the arguments advanced by the ethologist and roboticist David McFarland (1989), "Communication is the only behavior that requires an organism to self-monitor its own control system." Organisms can very effectively control themselves by a collection of competing but "myopic" task controllers, each activated by a condition (hunger or some other need, sensed opportunity, built-in priority ranking, and so on). When a controller's condition outweighs the conditions of the currently active task controller, it interrupts it and takes charge temporarily. (The "pandemonium model" by Oliver Selfridge [1959] is the ancestor of many later models.) Goals are represented only tacitly, in the feedback loops that guide each task controller, but without any global or higher level representation. Evolution will tend to optimize the interrupt dynamics of these modules, and nobody's the wiser. That is, there doesn't have to be anybody home to be wiser!

Communication, McFarland claims, is the behavioral innovation which changes all that. Communication requires a central clearing house of sorts in order to buffer the organism from revealing too much about its current state to competitive organisms. As Dawkins and Krebs (1978) showed, in order to understand the evolution of communication we need to see it as grounded in manipulation rather than as purely cooperative behavior. An organism that has no poker face, that "communicates state" directly to all hearers, is a sitting duck, and will soon be extinct (von Neumann and Morgenstern 1944). What must evolve to prevent this exposure is a private, proprietary communication-control buffer that creates opportunities for *guided* deception—and, coincidentally, opportunities for self-deception (Trivers 1985)—by creating, for the first time in the evolution of nervous systems, explicit and more globally accessible representations of its current state, representations that are detachable from the tasks they represent, so that deceptive behaviors can be formulated and controlled without interfering with the control of other behaviors.

It is important to realize that by communication, McFarland does not mean specifically *linguistic* communication (which is ours alone), but *strategic* communication, which opens up the crucial space between one's actual goals and intentions and the goals and intentions one attempts to communicate to an audience. There is no doubt that many species are genetically equipped with relatively simple communication behaviors (Hauser 1996), such as stotting, alarm calls, and territorial marking and defense. Stereotypical deception, such as bluffing in an aggressive encounter, is common, but a more productive and versatile talent for deception requires McFarland's private workspace. For a century and more philosophers have stressed the "privacy" of our inner thoughts, but seldom have they bothered to ask why this is such a good design feature. (An occupational blindness of many philosophers: taking the manifest image as simply *given* and never asking what it might have been given to us *for*.)

How did our manifest image become manifest to us?

Here is yet another strange inversion: this practice of sharing information in communicative actions with others, giving and demanding reasons, is what creates our personal user-illusions. All organisms, from single cells to elephants, have a rudimentary "sense of self." The amoeba adroitly keeps the bad stuff out and lets the good stuff in, protecting its vital boundaries. The lobster "knows enough" not to rip off and eat its own appendages. The free-floating rationales of the behavior of all *organisms* are *organized* around self-protection. In our case, the behaviors include a host of covert thinking behaviors we pick up in the course of enculturation, a process requiring lots of overt interaction with conspecifics. Practice makes perfect, and the sharpening and extending of these talents depends on a heightened level of mutual accessibility. The frolicking of puppies and bear cubs hones their abilities to perceive and anticipate each other's moves and to perceive and modulate their own actions and

reactions, a fine preparation for the more serious activities of adulthood. We humans need to develop a similar rapport with each other as we learn to communicate, and this requires perceiving *ourselves* in the execution of these behaviors. This is what gives us a less rudimentary, more "selfy" sense of self. We need to keep track of not only which limbs are ours and what we're doing with them but also which thoughts are ours and whether we should share them with others. We can give this strange idea an almost paradoxical spin: it is like something to be you *because* you have been enabled to tell us—or refrain from telling us—what it's like to be you!

When we evolved into an *us*, a communicating community of organisms that can compare notes, we became the beneficiaries of a system of user-illusions that rendered *versions* of our cognitive processes—otherwise as imperceptible as our metabolic processes—accessible to *us* for purposes of communication. McFarland is far from being the first to express the idea that explaining ourselves to others is the novel activity that generates the R&D that creates the architecture of human consciousness, as the epigraphs at the beginning of this section show. The idea purports to provide the basis for the long-sought explanation of the evolution of distinctively human consciousness. If it is mistaken, then at least it provides a model for what a successful account would have to accomplish. A number of thinkers have recently been homing in on related and congenial ideas: among them Douglas Hofstadter's "active symbols" (1979, 1982b, 1985 [esp. pp. 646ff], 2007), and three books in 2013, by psychologist Matthew Lieberman, neuroscientist Michael Graziano, and philosopher of cognitive science Radu Bogdan.

The evolution of memes provides the conditions for the evolution of a user interface that renders the memes "visible" to the "self" which (or who) communicates with others, the self as a *center of narrative gravity* (Dennett 1991), the author of both words and deeds. If joint attention to a shared topic is required (see the discussion of Tomasello in chapter 12), there have to be things—affordances—that both the first and the second person can attend to, and *this is what makes our manifest image manifest to us.* If we didn't have to

be able to talk to each other about our current thoughts and projects, and our memories of how things were, and so forth, our brains wouldn't waste the time, energy, and gray matter on an edited digest of current activities, which is what our stream of consciousness is. The self who has limited access to what is happening in its brain is well designed to entertain new memes, spread old memes, and compare notes with others. And what is this self? Not a dedicated portion of neural circuitry but rather like the end-user of an operating system. As Daniel Wegner put it, in his groundbreaking book *The Illusion of Conscious Will* (2002), "We can't possibly know (let alone keep track of) the tremendous number of mechanical influences on our behavior because we inhabit an extraordinarily complicated machine" (p. 27). Isn't it remarkable how easily we can follow Wegner into this apparently dualistic vision of ourselves as distinct *occupants* of our bodies! These machines "we inhabit" simplify things for our benefit: "The experience of will, then, is the way our minds portray their operations to us, not their actual operation" (p. 96).

Curiously, then, our *first-person* point of view of our own minds is not so different from our *second-person* point of view of others' minds: we don't see, or hear, or feel, the complicated neural machinery churning away in our brains but have to settle for an interpreted, digested version, a user-illusion that is so familiar to us that we take it not just for reality but also for the most indubitable and intimately known reality of all. That's what it is like to be us. We learn about others from hearing or reading what they say to us, and that's how we learn about ourselves as well. This is not a new idea, but keeps being rediscovered apparently. The great neurologist John Hughlings Jackson once said, "We speak, not only to tell others what we think, but to tell ourselves what we think" (1915). I, and many others, have misquoted the novelist and critic E. M. Forster as saying, "How do I know what I think until I see what I say?" Although Forster does have a version of this line in his book of criticism *Aspects of the Novel* (1927), he means it sarcastically and alludes to an earlier anecdote from which he draws it. This viral mutation of the Forster meme has spread widely, according to R. J. Heeks (2013), who shows

that the quote in context is meant to disparage the writing method of André Gide:

> Another distinguished critic has agreed with Gide—that old lady in the anecdote who has accused her nieces of being illogical. For some time she could not be brought to understand what logic was, and when she grasped its true nature she was not so much angry as contemptuous. "Logic! Good gracious! What rubbish!" she exclaimed. "How can I tell you what I think till I see what I say?" Her nieces, educated young women, thought that she was passée; she was really more up-to-date than they were. (Forster 1927, p. 71)

I am happy to set the record straight—or straighter, since I can find no trace of the anecdote about the lady and her nieces—but want to suggest that Forster walked by an important, if counterintuitive, possibility without noticing it. Our access to our own thinking, and especially to the causation and dynamics of its subpersonal parts, is really no better than our access to our digestive processes; we have to rely on the rather narrow and heavily edited channel that responds to our incessant curiosity with user-friendly deliverances, only one step closer to the real me than the access to the real me that is enjoyed by my family and friends. Once again, consciousness is not just talking to yourself; it includes all the varieties of self-stimulation and reflection we have acquired and honed throughout our waking lives. These are not just things that happen in our brains; they are behaviors that we engage in (Humphrey 2000, 2006, 2011), some "instinctively" (thanks to genetic evolution) and the rest acquired (thanks to cultural evolution and transmission, and individual self-exploration).

Why do we experience things the way we do?

If, as Wegner puts it, "our minds *portray* [my emphasis] their operations to us," if (as I have just said) your individual consciousness is

rather like the user-illusion on your computer screen, doesn't this imply that there *is* a Cartesian Theater after all, where this portrayal happens, where the show goes on, rather like the show you perceive on the desktop? No, but explaining what to put in place of the Cartesian Theater will take some stretching of the imagination.

We can list the properties of the tokens on the computer desktop: blue rectangular "files"; a black, arrow-shaped cursor; a yellow-highlighted word in black Times New Roman 12-point font; and so forth. What are the corresponding properties of these internal, re-identifiable private tokens in our brains? We don't know—yet. In chapter 9 we considered the way that bare meanings, with no words yet attached, could occupy our attention in consciousness, especially in the tip-of-the-tongue phenomenon. These are genuine tokens, tokens of memes or of sensation-types we are born with, or of other remembered affordances that can be recognized and re-identified even if they have no names (yet). Close your eyes and imagine a blue capital A. Done? You just created a token in your brain, but we can be sure that it isn't blue, any more than the tokens of "o" that occur in a word-processing file are round. The tokenings occur in the activity of neural circuits, and they have an important role to play in directing attention, arousing associated tokens, and modulating many cognitive activities. They contribute to directing such fundamental actions as saccadic eye movements and to initiating such higher level actions as awakening dozens of apps—memes—that are, as always, bent on getting new tokens of themselves—offspring—into the arena. Look:

tigr strp

The visual experience I just provided probably awakened the words "tiger" and "stripe" in your mind, and probably—did you notice?—these tokens had a specifically "auditory" cast, the long *i* in both words being somewhat highlighted in contrast to the almost unpronounceable visual stimuli that awakened them. These words then populated your neural workspace with *represen-*

tations of black and orange stripes that were not themselves black or orange, of course. (Were you actually aware of orange and black tiger stripes in your visual imagination? Maybe not, because the activation wasn't quite strong enough in your case, but you can be quite sure that the subpersonal [and subconscious] tokens were activated, since they would "prime" your answers to other questions in an experimental setting.)

All this subpersonal, neural-level activity is where the actual causal interactions happen that provide your cognitive powers, but all "you" have access to is the results. You can't tell by introspection how "tigr" got succeeded by "tiger" which then got succeeded by a "mental image" of a tiger, "focusing" on its stripes. When you attempt to tell us about what is happening in your experience, you ineluctably slide into a metaphorical idiom simply because you have no deeper, truer, more accurate knowledge of what was going *inside* you. You cushion your ignorance with a false—but deeply tempting—model: you simply reproduce, with some hand waving and apologies, your everyday model of how you know about what is going on *outside* you.

Here's how it happens. Let's start by reminding ourselves of something familiar and well understood: we send a reporter to observe some part of the external world—a nearby house, let's say—and report back to us by cell phone. Our phone rings, and when we answer, he tells us that there are four windows on the front side of the house, and when we ask him how he knows; he responds, "Because I'm looking right at them; I see them plain as day!" We typically don't think of then asking him how the fact that he sees them plain as day explains how he knows this fact. Seeing is believing, or something like that. We tacitly take the unknown pathways between his open eyes and speaking lips to be secure, just like the requisite activity in the pathways in the cell towers between his phone and ours. We're not curious on the occasion about how telephones work; we take them for granted. We also don't scratch our heads in bafflement over how he can just open his eyes and then answer questions with high reliability about what is positioned in

front of him in the light, because we all can do it (those of us who are not blind). How does it work? We don't know and are not usually curious about it.

When we do get curious, and ask him to describe, not the outside world, but his *subjective experience* of the outside world, his *inside* world, we put him on the spot, asking him to perform a rather unnatural act, and the results—unless he is a practiced *introspector* of one school or another—tend to be disappointing: "I dunno. I look out and see the house. That is, I think I see a house; there's a house shaped thing that seems to be about fifty yards away, and it has four window-looking things on it . . . and if I close my eyes and reopen them, it's still there and . . ."

The relative accessibility and familiarity of the outer part of the process of telling people what we can see—we know our eyes have to be open, and focused, and we have to attend, and there has to be light—conceals from us the utter blank (from the perspective of introspection or simple self-examination) of the rest of the process. We have no more privileged access to that part of the process than we do to the complicated processes that maintain the connectivity between our reporter's cell phone and ours.

How do you know there's a tree next to the house?
Well, there it is, and I can see that it looks just like a tree!
How do you know it looks like a tree?
Well, I just do!
Do you compare what it looks like with many other things in the world before settling upon the idea that it's a tree?
Not consciously.
Is it labeled "tree"?
No, I don't need to "see" a label; besides, if there were a label, I'd have to read it, and know that it was the label for the thing it was on. I just know it's a tree.

Imagine now that you could just spread your toes and thereby come to have breathtakingly accurate convictions about what was

currently happening in Chicago. And imagine not being curious about how that was possible.

How do you do it?

Not a clue, but it works, doesn't it? If I close my toes tight, I can no longer do it, and when I open them up again, whatever strikes my curiosity about current events in Chicago is instantly available to me. I just know.

What is it like?

Well, it's sort of like seeing and hearing, as if I were watching a remote feed television, but yet it's not quite like that. I just find all my Chicago-curiosity satisfied effortlessly.

Explanation has to stop somewhere, and at the personal level it stops here, with brute abilities couched in the familiar mentalistic language of knowing and seeing, noticing and recognizing, and the like. The problem with the first-person point of view is that it is anchored in the manifest image, not the scientific image, and cannot avail itself of the resources of the scientific image. The standard presumption is that when we challenge a reporter, "I know because I can see it" is an acceptably complete reply, but when we import the same presumption into the case where a subject is reporting on mental imagery or memory (or the imagined opened-toed Chicago clairvoyance), for instance, we create an artifact. What our questions *directly* create, or provoke, are answers, as in the dialogues above. What they indirectly create are ideologies based on those answers. You can ask yourself what your subjective experience is, and see what you want to say. Then you can decide to endorse your own declaration, to believe it, and then pursue the implications of that creed. You can do this by talking aloud to yourself, talking silently to yourself, or "just thinking" to yourself about what you are currently experiencing.

That is the extent of your access to your own experience, and it does not differ much from the access another person can have to those experiences—*your* experiences—if you decide to go public

with your account. Your convictions are no doubt reliable but not infallible. Another person could help you test them and perhaps get you to adjust them in the face of further experiences. This is the way to study consciousness scientifically, and I have given it an ungainly but accurate name: *heterophenomenology*, the phenomenology of the *other* person's experience, as contrasted with *autophenomenology*, the phenomenology of one's own experience. There is a long-standing tradition to the effect that somehow autophenomenology is a more intimate, more authentic, more direct way of getting at the objects of experience, that adopting the "first-person point of view" is the key strategic move in any promising study of consciousness, but that is itself a delusion. Heterophenomenology is more accurate, more reliable, less vulnerable to illusion than autophenomenology, once you control for lying and other forms of noncooperation with the investigation, and you can get a better catalogue of *your own* experience by subjecting yourself to all the experimental circumstances in which consciousness is studied. You can be *shown* features of your own experience of which you had no inkling, both unimagined absences and weaknesses and surprising abilities you didn't know you have.

Collaborating with other investigators on the study of your own consciousness (adopting, if you like, the "second-person point of view") is the way to take consciousness, as a phenomenon, as seriously as it can be taken. Insisting, in resistance to this, that you know more about your own consciousness just because it's yours, is lapsing into dogma. By shielding your precious experience from probing, you perpetuate myths that have outlived their utility.

We ask a subject to tell us how many windows there were in his bedroom in the house he grew up in, and he closes his eyes for a moment and replies "two." We ask: How do you know? "Because I just 'looked' . . . and I 'saw' them!" He didn't literally look, of course. His eyes were closed (or were staring unfocused into the middle distance). The "eyes part" of the seeing process wasn't engaged, but a lot of the rest of the vision process was—the part that we normally don't question. It's sort of like seeing and sort of not like seeing,

but just how this works is not really accessible to folk-psychological exploration, to introspection, or self-manipulation. When we confront this familiar vacuum, there is an almost irresistible temptation to postulate a surrogate world—a mental image—to stand in for the part of the real world that a reporter observes. And we can be sure of the existence of such a surrogate world in one strained sense: there has to be *something* in there—something in the neural activity—that reliably and robustly maintains lots of information on the topic in question since we can readily confirm the fact that information can be extracted from "it" almost as reliably as from real-world observation of a thing out there. The "recollected image" of the house has a certain richness and accuracy that can be checked, and its limits gauged. These limits give us important clues about how the information is actually embodied in the brain, and we mustn't just jump to the conclusion that it is embodied, as it seems to be, in an image that can be consulted.[99]

From this perspective, our utter inability to say what we're doing when we do what we call framing mental images is not so surprising. Aside from the peripheral parts about what we're doing with our eyes, we are just as unable to say what we're doing in a case of seeing the external world. We just look and learn, and that's all we know. Consider the subpersonal processes of normal vision, and note that at some point they have to account for all the things we can do, thanks to having our eyes open: we can pick blueberries, hit baseballs, recognize landmarks, navigate through novel terrain, and read, for instance. Also, thanks to these processes, our internal cortical states suffice to guide our speaking-subsystem in the framing of descriptive speech acts. We are making steady progress on this subpersonal story, even if large parts of it remain quite baffling today. We can be confident that there will be a subpersonal story that goes all the way from the eyeballs to oral reports (among other things), and in that story there will *not* be a second

99 This is where the experimental and theoretical work on mental imagery by Roger Shepard, Stephen Kosslyn, Zenon Pylyshyn, and many others comes into play.

presentation process with an ego (a self, a boss, an inner witness) observing an inner screen and then composing a report. As I never tire of saying, *all* the work done by the imagined homunculus in the Cartesian Theater has to be broken up and distributed around (in space and time) to lesser agencies in the brain.

Well then, let's try to break up the self into some plausible parts. What diminutions, what truncations, of the observing reporter might do the trick? Perhaps "agents" that were full of convictions but clueless about how they came by them—rather like oracles, perhaps, beset with judgments but with nothing to tell us (or themselves, of course) about how they arrived at that state of conviction. Ray Jackendoff addressed this issue some years ago (1996) and offered a useful prospect, recently elaborated on by Bryce Huebner and me, introducing the concept of a subpersonal *blurt* (2009):

> The key insight is that a module "dumbly, obsessively converts thoughts into linguistic form and vice versa" (Jackendoff 1996). Schematically, a conceptualized thought triggers the production of a linguistic representation that approximates the content of that thought, yielding a reflexive *blurt*. Such linguistic *blurts* are proto-speech acts, issuing subpersonally, not yet from or by the person, and they are either sent to exogenous broadcast systems (where they become the raw material for personal speech acts), or are endogenously broadcast to language comprehension systems which feed directly to the mind-reading system. Here, *blurts* are tested to see whether they should be uttered overtly, as the mind-reading system accesses the content of the *blurt* and reflexively generates a belief that approximates the content of that *blurt*. Systems dedicated to belief fixation are then recruited, beliefs are updated, and the *blurt* is accepted or rejected, and the process repeats. (Huebner and Dennett 2009, p. 149)

This is all very impressionistic and in need of further details, but a similar idea of "narrative" making has still more recently been

developed by Gustav Markkula (2015), who argues persuasively that the human activity of (roughly) asking and telling ourselves what it is like to be us, creates artifacts of imagination that we take to be the "qualia" so beloved by philosophers of consciousness who yearn to reinstate dualism as a serious theory of the mind.

Hume's strange inversion of reasoning

But still, comes the objection, why does it have to be like anything to see, hear, and smell? Why does there *seem* to be an inner theater with a multimedia show going on whenever we're awake? Even if we grant that there must be a subpersonal story, in the scientific image, that can satisfactorily explain all the behaviors and emotional responses, the decisions and verbal reports I make, it must leave *me* out of the story! And putting me and my qualia back in the world, not leaving them out, is a task still to be done. The best response I know to this challenge is what I call Hume's strange inversion of reasoning, because he articulated a prescient account of one case—our experience of causation itself—long before Darwin and Turing discovered their inversions. There are complications and controversies aplenty about Hume's theory of causation, and some of his account, influential as it has been for several centuries, is largely abandoned today, but one central idea shines through and provides an important insight about the relation between the manifest and scientific images, and about the nature of our conscious experience in general, not just our experience of causation.

We seem to *see* and *hear* and *feel* causation every day, Hume notes, as when we see a brick shatter a window or hear a bell ring on being struck, but all we ever directly experience, Hume insists, is sequence: A *followed by* B, not A *causing* B. If Hume were wrong about this, animated cartoons would be impossible: when representing Bugs Bunny chomping on a carrot, the animators would have to add not just a sound track, with a loud *crunch* synchronized with the bite, but some kind of *cause track* that would *show us directly*

somehow that Bugs's teeth closing actually causes, and doesn't just immediately precede, the disappearance of half the carrot and the crunching noise we hear. There is no such need, of course. Mere succession of one frame of film by the next is all it takes to create the *impression* of causation. But, Hume notes, the impression of causation we experience comes from inside, not outside; it is itself an effect of a habit of expectation that has been engrained in us over many waking hours. (Hume insisted that these expectation habits were all learned, acquired during normal infancy, but contemporary research strongly suggests that we are born with a sort of automatic causal sense, like a reflex, that is ready to "see" causation whenever our senses are confronted by the right kind of sequence of stimuli.) Seeing A, we are *wired* to expect B, and then when B happens—this is Hume's master stroke—we *misattribute* our perceptual reaction to some external cause that we are somehow directly experiencing. (We think we *see* Bugs Bunny's cartoon teeth *cause* the lopping off of the carrot.) In fact, we are succumbing to a benign user-illusion, misinterpreting our fulfilled expectation of an ensuing B as somehow coming from the outer world. This is, as Hume says, a special case of the mind's "great propensity to spread itself on external objects" (1739, I:xiv). The "customary transition" in our minds is the source of our sense of causation, a quality of "perceptions, not of objects," and, as he notes, "the contrary notion is so riveted in the mind" that it is hard to dislodge. It survives to this day in the typically unexamined assumption that all perceptual representations must be flowing inbound from outside.

Here are a few other folk convictions that need Hume's strange inversion: sweetness is an "intrinsic" property of sugar and honey, which causes us to like them; observed intrinsic sexiness is what causes our lust; it was the funniness out there in the joke that caused us to laugh (Hurley, Dennett, Adams 2011). Oversimplifying somewhat, in these instances the causes and effects in the manifest image have become inverted in the scientific image. You can't find intrinsic sweetness by studying the molecular structure of

glucose; look instead to the details in the brains of sweetness seekers. It is how our brains respond that causes "us" (in the manifest image) to "project" an illusory property into the (manifest) world. There are structural, chemical properties of glucose—mimicked in saccharine and other artificial sweeteners—that cause the sweetness response in our nervous systems, but "the intrinsic, subjective sweetness I enjoy" is not an internal recreation or model of these chemical properties, nor is it a very special property in our nonphysical minds that we use to decorate the perceptible things out there in the world. It is no property at all; it is a benign illusion. Our brains have tricked us into having the conviction, making the judgment, that there seems to be an intrinsically wonderful but otherwise undescribable property in some edible things: sweetness. We can recognize it, recall it, dream about it, but we can't describe it; it is ineffable and unanalyzable.

There is no more familiar and appealing verb than "project" to describe this effect, but of course everybody knows it is only metaphorical; colors aren't literally projected (as if from a slide projector) out onto the front surfaces of (colorless) objects, any more than the idea of causation is somehow beamed out onto the point of impact between billiard balls. If we use the shorthand term "projection" to try to talk, metaphorically, about the mismatch between manifest and scientific image here, what is the true long story? What is literally going on in the scientific image? A large part of the answer emerges, I propose, from the predictive coding perspective we explored briefly in chapter 8 (How do brains pick up affordances?).

Here is where Bayesian expectations can play an iterated role: our ontology (in the elevator sense) does a close-to-optimal job of cataloguing the things in the world that matter to the behavior our brains have to control. Hierarchical Bayesian predictions accomplish this, generating affordances galore: we expect solid objects to have backs that will come into view as we walk around them, doors to open, stairs to afford climbing, cups to hold liquid, and so forth. But among the things in our *Umwelt* that matter to our well-being are *ourselves*! We ought to have good Bayesian expectations about

what we will do next, what we will think next, and what we will *expect* next! And we do. Here's an example:

Think of the cuteness of babies. It is not, of course, an "intrinsic" property of babies, though it seems to be. What you "project" onto the baby is in fact your manifold of "felt" dispositions to cuddle, protect, nurture, kiss, coo over, . . . that little cutie-pie. It's not just that when your cuteness detector (based on facial proportions, etc.) fires, you have urges to nurture and protect; you *expect* to have those very urges, and that manifold of expectations just *is* the "projection" onto the baby of the property of cuteness. When we expect to see a baby in the crib, we also expect to "find it cute"—that is, we *expect* to *expect* to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction-error signals is interpreted by our brains as confirmation that, indeed, the thing in the world we are interacting with really has the properties we expected it to have. Cuteness as a property *passes the Bayesian test for being an objective structural part of the world we live in*, and that is all that needs to happen. *Any further "projection" process would be redundant.* What is special about properties like sweetness and cuteness is that their perception depends on particularities of the nervous systems that have evolved to make much of them. They have a biased or privileged role in the modulation of our control systems—we care about them, in short.

Here we must be very careful not to confuse two independent claims. The properties of sweetness and cuteness depend on features of our nervous systems and hence are in that limited sense subjective, but that must not be taken to mean that sweetness, say, is an *intrinsic* (subjective) property of conscious experience! Hume's strange inversion is wonderful but incomplete: when he spoke of the mind's "great propensity to spread itself on external objects," this should be seen not as a stopping point but as a stepping-stone to a further inversion. Hume's image brilliantly conjures up the curious vision of the mind painting the external world with the proprietary ("intrinsic") hues properly worn by the mind's internal items—impressions and ideas, in his vocabulary. But there is no such paint

(which is why I once dubbed it "figment"). We need to push Hume's inversion a little harder and show that the icons of the user-illusion of our minds, unlike the user-illusion of our computers, don't need to be rendered on a screen.

A red stripe as an intentional object

One more example should at least clarify my point, if not succeed in persuading everybody—as Hume says, the contrary notion is so riveted in our minds. Look steadily at the white cross in the top panel of figure 14.1 of the color insert (following p. 238) for about ten seconds, and then switch your gaze to the white cross in the bottom panel. What do you see?

"I see an American flag, red white, and blue."
"Do you see a red stripe at the top on the right?" (Do the experiment again.)
"Yes, of course. There is a fuzzy, faint red stripe to the right of the field of blue with the stars."

But think: there are no red stripes on the page, on your retina, or in your brain. In fact, there is no red stripe anywhere. It just seems to you that there is a red stripe. Your brain "projects" a nonexistent red stripe onto the world. (It is important that the illusory stripe doesn't appear to you to be in your head; it appears to be on the page, as if projected there by a video projector in the middle of your forehead.) The phenomenon in you that is responsible for this is *not* a red stripe. It is a representation of a red stripe in some neural system of representation that we haven't yet precisely located and don't yet know how to decode, but we can be quite sure it is neither red nor a stripe. You don't know exactly what causes you to seem to see a red stripe out in the world, so you are tempted to lapse into Humean misattribution: you misinterpret your sense (judgment, conviction, belief, inclination) that you're seeing a red stripe as arising from a subjective property (a *quale*, in the jargon of philosophy)

that is the *source* of your judgment, when in fact, that is just about backward. It is your *ability to describe* "the red stripe," your judgment, your willingness to make the assertions you just made, and your emotional reactions (if any) to "the red stripe" that is the source of your conviction that there *is* a subjective red stripe.

This is an instance of a kind of mistake that has been much examined in other quarters of philosophy: mistaking the *intentional object* of a belief for its *cause*. Normally, when you're not being tricked by your senses, or by tricksters, when you believe in something (in the existence of something with certain features in your vicinity, for instance), it is because just such a thing with just those features has caused you to believe in it, by stimulating your sense organs. You believe in the apple in your right hand *because* that very apple has *caused* you to believe in its existence, reflecting light into your eyes, and exerting a downward force on your palm. In this sort of normal case, we can say, carefully setting aside quibbles, that the apple, the intentional object of your belief is also the (primary, or salient) cause of your belief. But there are well-known abnormal cases: mirages, optical illusions, hallucinations, and complementary color afterimages—and pranks. Suppose a group of us decide to play a mean trick on Otto: we concoct a phony person, Dan Quale, and proceed to plant e-mails, text messages, and birthday cards to Otto from Dan Quale along with footprints, phone calls, and very close—but not too close—encounters that Otto is maneuvered into having with the elusive (in fact fictitious) Quale. Pretty soon Otto believes Dan Quale is a real person, with a fairly detailed recent biographical trail, a voice, a stature, and a lot more. That Dan Quale is the intentional object of a manifold of beliefs Otto has. The beliefs are all *about* Dan Quale, even though Dan Quale does not exist. Lots of other people exist, and lots of footprints and e-mails and all the rest, but not Dan Quale. Otto's beliefs about Dan Quale have a panoply of semi-organized causes, none of which is a person named Dan Quale. But Otto doesn't know this. He is really quite sure that Dan Quale exists—he's seen him, talked to him on the phone, has letters from him, and so forth. He wants to meet Dan Quale. Whom does he

want to meet? Not any of the pranksters; he knows them all and has no particular desire to see any of them again. Dan Quale doesn't exist but is the intentional object of Otto's quest—just like Ponce de Leon searching for the Fountain of Youth. Ponce *had an idea in his mind* of the Fountain of Youth (we might say, loosely speaking), but that mental state was not the object of his quest. He already had it! He wasn't seeking an idea; he was seeking a fountain. And Otto isn't seeking his mental states about Dan Quale. He's seeking a man, driven in that search by his mental states.

Now apply the same analysis to the red stripe. If you didn't know about complementary color afterimages, you might well be naïvely confident that there really was a red stripe, visible to others from the same vantage point, in the "external" world. If *that* is what you believe, then the intentional object of your belief does not exist, and among its causes are a green striped flag picture and a lot of neural events in your visual cortex, none of which are red or even appear to be red. You are not that naïve and know very well that no such external red stripe exists, and this may mislead you to the conjecture (or, in many cases, the adamant conviction) that you couldn't be wrong—*there is* a "subjective" red stripe in your mind. You *see* it! Well, you sorta see it. In support of this theoretical postulation, you may ask: How *could* I be having an experience of a horizontal red stripe unless something somewhere is red and horizontal? The somewhat rude answer to this rhetorical question is, "Easy. If you can't conceive of this, try harder."

This moment marks the birth of qualia, an artifact of bad theorizing. The intentional object of your beliefs is not in doubt: you believe with all your heart and soul—not that there is a red stripe *out there* but—that there is a red stripe *in here* (something with the qualia of a red stripe): after all, you can "look at it," "concentrate on it," "recall it," "enjoy it," "compare it with other such things in memory." Qualia are supposed to be the somehow internal, subjective properties that we are acquainted with more directly, when we are slightly less directly acquainted with their normal external causes—real red stripes, and so on in the world. When you make

this move, you are positing an *internal cause* that has the same properties as the *intentional objects* that normally cause your perceptual beliefs—except that these are private, subjective versions, somehow, of the public, objective properties of redness and so forth. But once you realize that the intentional objects of mistaken beliefs simply don't exist, anywhere, you have no need in your theory or conjecture for weird internal something-or-others with mysterious properties. Dan Quale, the intentional object of Otto's Dan-Quale-beliefs, isn't made of ectoplasm or fictoplasm or anything. Neither is Santa Claus or Sherlock Holmes. So when you seem to see a red stripe when there is no red stripe in the world as its source, there need be no *other thing* (made of red figment) that is the "real seeming" you take yourself to be experiencing.

What is there in its place? What *does* explain your conviction that there is a red stripe? The presence in your brain of *something*—yes, of course there has to be something in your brain that is responsible—but it is something in the medium of neural spike train activity, not some other medium: a remarkably salient, information-rich, subpersonal state, a token of a red stripe representation that is no more red nor striped than those neural word tokens described in Chapter 9 were loud or soft (or red or black). That is the cause of your belief in the red stripe, but it is not the intentional object of your belief (because it isn't red or striped).

But even if this is a *possible* explanation of all my subjective states, how do we know there isn't a qualia medium somehow in our minds, if not in our brains? How do we know that the "naïve" theory is mistaken? Suppose that instead of answering the rhetorical question rudely, we surrender to it, for the moment, and see what follows. Let's suppose then that there *is* a subjective property of some kind that "explains" your current introspective convictions and abilities. Let's suppose, that is, that when you experience what seems to be a horizontal red stripe, there really is, somewhere, a horizontal-shaped red quale (whatever that is) and it is somehow the cause or source of your conviction that you are experiencing a horizontal red stripe, and that this *rendering* in some unknown *medium* is caused

or triggered by the confirmation (the absence of disconfirmation) of all the expectations generated by the normal operation of your visual system. Just to make the supposition as clear as possible, here is a somewhat expanded version of the purported explanation of the red afterimage effect:

> Fixating on the real green stripes in front of you for a few seconds fatigues the relevant neural circuits in the complementary color system, which then generate a false signal (red, not green), which does not get disconfirmed so long as the fatigue lasts, so somewhere fairly high in the process betwixt retina and, um . . . the philosophical conviction center, a red stripe-shaped quale is rendered, and it is the appreciation of this quale that grounds, fuels, informs, causes, underwrites the philosophical conviction that right now you are enjoying a stripe-shaped red quale.

This spells out the idea behind the rhetorical question: We need *something* like this—don't we?—to *explain* the undeniable fact that it sure seems to you there's a red stripe right now. You're not just saying this (the way a robot might, if programed to be a model of complementary color afterimages); you believe it with all your heart and soul.

Fine. So now we have qualia installed in our sketchy model of the process. What next? Something would have to *have access* to the rendering in that medium (otherwise, the rendered qualia would be wasted, unwitnessed, and unappreciated, like a beautiful painting locked in an empty room). Call whatever it is that has this access the inner observer. Now what do you suppose an appropriate reaction to this rendering by this inner observer would be? What else but the judgment that there sure seems to be a red stripe out there, part of an apparent American flag? But that conclusion had already been arrived at in the course of the nondisconfirmed expectations. A red stripe in a particular location in visual space had already been identified by the system; that conclusion was the information that

informed the inner rendering (the way a bitmap informs the rendering of colors on your computer screen). The postulation of qualia is just doubling up the cognitive work to be done. There is no more work (or play) for consciousness to do.

This is the importance of always asking what I have called the Hard Question (1991, p. 255): *And then what happens?* Many theorists of consciousness stop with half a theory. If you want a whole theory of consciousness, this is the question you must ask and answer after you have delivered some item "to consciousness" (whatever you take arrival in consciousness to amount to). If instead you just stop there and declare victory, you've burdened the Subject or Self with the task of reacting, of doing something with the delivery, and you've left that task unanalyzed. If the answer you give to the Hard Question ominously echoes the answer you gave to the "easy" questions about how the pre-qualia part of the process works, you can conclude that you're running around in a circle. Stop. Reconsider.

Doggedly pursuing the idea that qualia are both the causes and the intentional objects (the *existing* intentional objects) of introspective beliefs leads to further artifactual fantasies, the most extravagant of which is the idea that unlike our knowledge of all other kinds of causation, our knowledge of mental causation is infallible and direct: we can't be wrong when we declare that our subjective beliefs about the elements of our conscious experience are caused by those very elements. We have "privileged access" to the *causes* or *sources* of our introspective convictions. No logical room for any tricksters intervening here! We *couldn't be* victimized by any illusions here! You might be a zombie, unwittingly taking yourself to have real consciousness with real qualia, but I *know* that I am not a zombie! No, you don't. The only support for that conviction is the vehemence of the conviction itself, and as soon as you allow the theoretical possibility that there *could* be zombies, you have to give up your papal authority about your own nonzombiehood. I cannot prove this, yet, but I can encourage would-be consciousness theorists to recognize the chasm created by this move and recognize that they can't have it both ways.

What is Cartesian gravity and why does it persist?

René Descartes wasn't the first great thinker to try to give an account of the human mind, but his vision, as presented in his *Discourse on Method* (1637) and *Meditations* (1641) was so vivid and compelling that it has strongly influenced all subsequent thought on the topic. His pioneering investigations of brain anatomy were intrepid and imaginative, but his tools and methods were unable to plumb more than a tiny fraction of the complexities he exposed, and the only available metaphors—wires and pulleys and fluids rushing through hoses—were too crude to furnish his imagination with a measure of the possibilities for a materialistic model of the brain as the mind. So he can hardly be faulted for jumping to the conclusion that the mind he knew so much about "from the inside" must be some *other* thing, a thinking thing (*res cogitans*) that was not material at all. He got off on the wrong foot, then, by taking the "first-person point of view" as his direct, and even infallible, epistemic access to consciousness, a step which anchored him in a user-illusion that systematically distorted the investigation from the outset. But what else could he do? Looking at brain tissue was preposterously uninformative compared with reflecting on his thoughts, the sensations and perceptions he enjoyed or abhorred, the plans he concocted, and the emotions that altered his moods.

Ever since, philosophers, psychologists, and other scientists have relied heavily on introspection as at least a bountiful source of hints (and problems), while postponing asking the question of how this marvelous treasure trove was possible. After all, it was "self-evident"; our conscious minds are filled with "ideas" and "sensations" and "emotions" of which we have "knowledge by *acquaintance*" that—most thinkers agreed—surpassed in intimacy and incorrigibility every other kind of knowledge. The primacy of "first-person experience" has been implicit in the practices of most investigators if not always a declared axiom. Sometimes it has even been upheld as fundamental methodological wisdom: John Searle (1980) lays it

down categorically: "Remember, in these discussions, always insist on the first person point of view. The first step in the operationalist sleight of hand occurs when we try to figure out how we would know what it would be like for others" (p. 451).[100] Indeed for many philosophers, the central problem has been not how to provide a scientific account of conscious experience, but how to penetrate the "veil of perception" and get from "in here" to the "external world," and Descartes's *Meditations* was the inaugural exploration of that way of thinking.

The price you pay for following Searle's advice is that you get all your *phenomena*, the events and things that have to be explained by your theory, through a channel designed not for scientific investigation but for handy, quick-and-dirty use in the rough and tumble of time-pressured life. You can learn a lot about how the brain does it—you can learn quite a lot about computers by always insisting on the desktop point of view, after all—but only if you remind yourself that your channel is systematically oversimplified and metaphorical, not literal. That means you must resist the alluring temptation to postulate a panoply of special subjective properties (typically called qualia) to which you (alone) have access. Those are fine items for our manifest image, but they must be "bracketed," as the phenomenologists say, when we turn to scientific explanation. Failure to appreciate this leads to an inflated list of things that need to be explained, featuring, preeminently, a Hard Problem that is nothing

100 Operationalism is the proposal by some logical positivists back in the 1920s that we don't know what a term means unless we can define an operation that we can use to determine when it applies to something. Some have declared that the Turing Test is to be taken as an *operationalist definition* of intelligence. The "operationalist sleight of hand" that Searle warns against is the claim that we really can't claim to know what consciousness is until we figure out how we can learn about the consciousness of others. Searle's alternative is itself a pretty clear case of operationalism: *If I want to know what consciousness is, my measurement operation is simple: I just look inside and whatever I see—that's consciousness!* It works for him, of course, but not for others.

but an artifact of the failure to recognize that evolution has given us a gift that sacrifices literal truth for utility.

Imagine asking for some advice and being advised, "Use your pancreas!" or "Use your liver!" You would have no idea what action to take. And when a teacher urges you to "use your brain" you'd be utterly stymied if you didn't interpret this as the directive to "use your *mind*," that thinking thing with which you are so intimately acquainted that it is hardly distinguishable from you, yourself. No wonder we are reluctant to see it as illusory; if *it* is illusory, so are *we*!

If we, our *selves*, were all "just" part of each other's user-illusions, wouldn't that imply that, really, life has no meaning? No. The manifest image that has been cobbled together by genetic evolutionary processes over billions of years, and by cultural evolutionary processes over thousands of years, is an extremely sophisticated system of helpful metaphorical renderings of the underlying reality uncovered in the scientific image. It is a user-illusion that we are so adept at using that we take it to be unvarnished reality, when in fact it has many coats of intervening interpretive varnish on it. The manifest image composes our *Umwelt*, the world we live in for almost all human purposes—aside from science. We learn about reality via the categories of colors, sounds, aromas, solid objects, sunsets and rainbows, people and their intentions, promises, threats, and assurances, institutions, and artifacts. We view our prospects, make our decisions, plan our lives, commit our futures in its terms, and this is what makes the manifest image *matter*—to us. It's life or death for us, and what else could make it matter more? Our own reflections on all this are necessarily couched in terms of meanings, or contents, the only readily usable "access" we have to what goes on between our ears and behind our eyes.

If Searle has stressed for years the importance of adopting a *first*-person point of view (what do *I* see? what is it like to be *me*?), a philosopher who early appreciated the "antiseptic virtue" of adopting the *third*-person point of view (what does *it* want? what are *they* conscious of?) is Jonathan Bennett, whose little monograph, *Rationality* (1964), set out to study human rationality indirectly by studying the (non)rationality of bees! By insisting on adopting the third-person point of view, and starting with a humble-but-oh-so-competent creature, Bennett minimizes the temptation to be taken in by the unavoidable practice of *identification by content* that is the hallmark of introspective methods.

This is what I mean: if you want to talk about your own mental states, you *must* identify them by their content: "Which idea? My idea of HORSE. Which sensation? My sensation of *white*." How else? There is no way you can identify your own mental states "from the inside" as, for instance, *concept J47* or *color-sensation 294*. By taking for granted the content of your mental states, by picking them out *by their content*, you sweep under the rug all the problems of indeterminacy or vagueness of content. Reading your own mind is too easy; reading the mind of a honeybee places the problems front and center. We won't have a complete science of consciousness until we can align our manifest-image identifications of mental states by their contents with scientific-image identifications of the subpersonal information structures and events that are causally responsible for generating the details of the user-illusion we take ourselves to operate in.

Here is another source of the staying power of the Cartesian point of view. By presupposing that we normal folks are rational and hence have understanding (not just competence), we tacitly endorse our everyday use of the intentional stance as not just practical and valuable but as also *the plain truth* about human minds. This puts us in distinguished company: we are intelligent designers, rather like the Intelligent Designer who designed us. We wouldn't want to give up that honor, would we? And so we normally give both ourselves and our fellow human beings more credit for the authorship of our creations, and more blame for our misdeeds, than would be warranted by an unvarnished view of the causation involved.

Besides—and here comes a big payoff—the Cartesian point of view fits nicely, it seems, with traditional ideas of free will and

moral responsibility. I recognized the penetration of this hunch some years ago when I tried to uncover the submerged grounds for resistance to any version of the account of consciousness sketched here, and I discovered that many cognitive scientists—not only lay-people—were reluctant even to *consider* such doctrines. After I had laid to rest a number of their objections, they would often eventually let the cat out of the bag: "But what about free will? Wouldn't a completely materialistic account of consciousness show that we can't be morally responsible?" No, it wouldn't, and here, in a nutshell, is why (I've had my say about that question in two books and many articles, so on this occasion I will be brief). The traditional view of free will, as a personal power somehow isolated from physical causation, is both incoherent and unnecessary as a grounds for moral responsibility and meaning. The scientists and philosophers who declare free will a fiction or illusion are right; it is part of the user-illusion of the manifest image. That puts it in the same category with colors, opportunities, dollars, promises, and love (to take a few valuable examples from a large set of affordances). If free will is an illusion then so are they, and for the same reason. This is not an illusion we should want to dismantle or erase; it's where we live, and we couldn't live the way we do without it. But when these scientists and philosophers go on to claim that their "discovery" of this (benign) illusion has important implications for the law, for whether or not we are responsible for our actions and creations, their arguments evaporate. Yes, we should shed the cruel trappings of retributivism, which holds people *absolutely* responsible (in the eyes of God) for their deeds; we should secure in its place a sane, practical, defensible system of morality and justice that still punishes when punishment is called for, but with a profoundly different framing or attitude. One can get a sense of this by asking yourself: If—because free will is an illusion—no one is ever responsible for what they do, should we abolish yellow and red cards in soccer, the penalty box in ice hockey, and all the other penalty systems in sports?

The phenomena of free will and moral responsibility, worthy items in the ontology of the human manifest image, survive robustly once we strip off some of the accrued magic of tradition and reground them in scientific reality. Regardless of whether I am right in my claim that the phenomena of free will and responsibility, properly reformed and understood, are defensible elements in our most serious ontology, we need to recognize how the fear that these important features of everyday life are doomed generates a powerful undercurrent of resistance that distorts the imaginations of people trying to figure out what human consciousness is.

Nicholas Humphrey is the writer who has most forcefully drawn attention to the unargued prejudice people tend to have in favor of "spiritual" accounts in his *Soul Searching: Human Nature and Supernatural Belief* (1995). As he shows, people tend to treat belief in the supernatural as not only excusable but also morally praiseworthy. Credulity is next to godliness. The human mind, many think, is the last bastion of what is sacred in this world, and to explain it would be to destroy it, so to be safe, we had better declare consciousness conveniently out of bounds to science. And as we have seen, there is tremendous emotional resistance to any considerations that might seem to cast doubt on the presumption that nonhuman animals—the grizzly bear, the puppy, the dolphin—have minds that are, if not *just* like ours, at least enough like it to provide them some moral standing, perhaps not moral responsibility but at least the right not to be ill treated.

For all these reasons, resisting the force of Cartesian gravity takes some strenuous exercises of the imagination, and we need to avoid overshooting in our resistance as well. We have to set aside some intuitions that seem *almost* indubitable and take seriously some suggestions that seem, at first, paradoxical. That's difficult, but science has shown us again and again how to do it. Even school children can shed pre-Copernican and pre-Galilean intuitions without flinching, and by the time they are teenagers they can get comfortable replacing some of their Newtonian intuitions with Einsteinian intuitions. Getting comfortable with quantum physics is still a work in progress—for me, I confess, in spite of much mental calisthenics.

Easier (for me, in any case) is embracing Darwin's strange inversion of reasoning, and Turing's, and Hume's. By offering a sketch of the *causes* of Cartesian gravity, I have tried to help the unpersuaded find a vantage point from which they can diagnose their own failures of imagination and overcome them.

Human consciousness is unlike all other varieties of animal consciousness in that it is a product in large part of cultural evolution, which installs a bounty of words and many other thinking tools in our brains, creating thereby a cognitive architecture unlike the "bottom-up" minds of animals. By supplying our minds with systems of representations, this architecture furnishes each of us with a perspective—a user-illusion—from which we have a limited, biased access to the workings of our brains, which we involuntarily misinterpret as a rendering (spread on the external world or on a private screen or stage) of both the world's external properties (colors, aromas, sounds, . . .) and many of our own internal responses (expectations satisfied, desires identified, etc.). The incessant torrent of self-probing and reflection that we engage in during waking life is what permits us, alone, to comprehend our competences and many of the reasons for the way the world is. Thanks to this infestation of culturally evolved symbiont information structures, our brains are empowered to be intelligent designers, of artifacts and of our own lives.

15

The Age of Post-Intelligent Design

What are the limits of our comprehension?

> If the brain were so simple we could understand it, we would be so simple we couldn't.
>
> —Emerson M. Pugh, *The Biological Origin of Human Values*

Human comprehension has been steadily growing since prehistoric times. For forty millennia and more, we have been living in the age of intelligent design—crafting pots, tools, weapons, clothes, dwellings and vehicles; composing music and poetry; creating art; inventing and refining agricultural practices; and organizing armies, with a mixture of dutiful obedience to tradition, heedless and opportunistic improvisation, and knowing, intentional, systematic R&D, irregularly punctuated with moments of "inspired" genius. We applaud intelligent design in all arenas, and aspire from infancy to achieve recognition for our creations. Among the artifacts we have created is the concept of God, the Intelligent Designer, in our own image. That's how much we value the intelligent designers in our societies.

We recognize the value of these fruits of our labors, and our laws and traditions have been designed to create an artificial