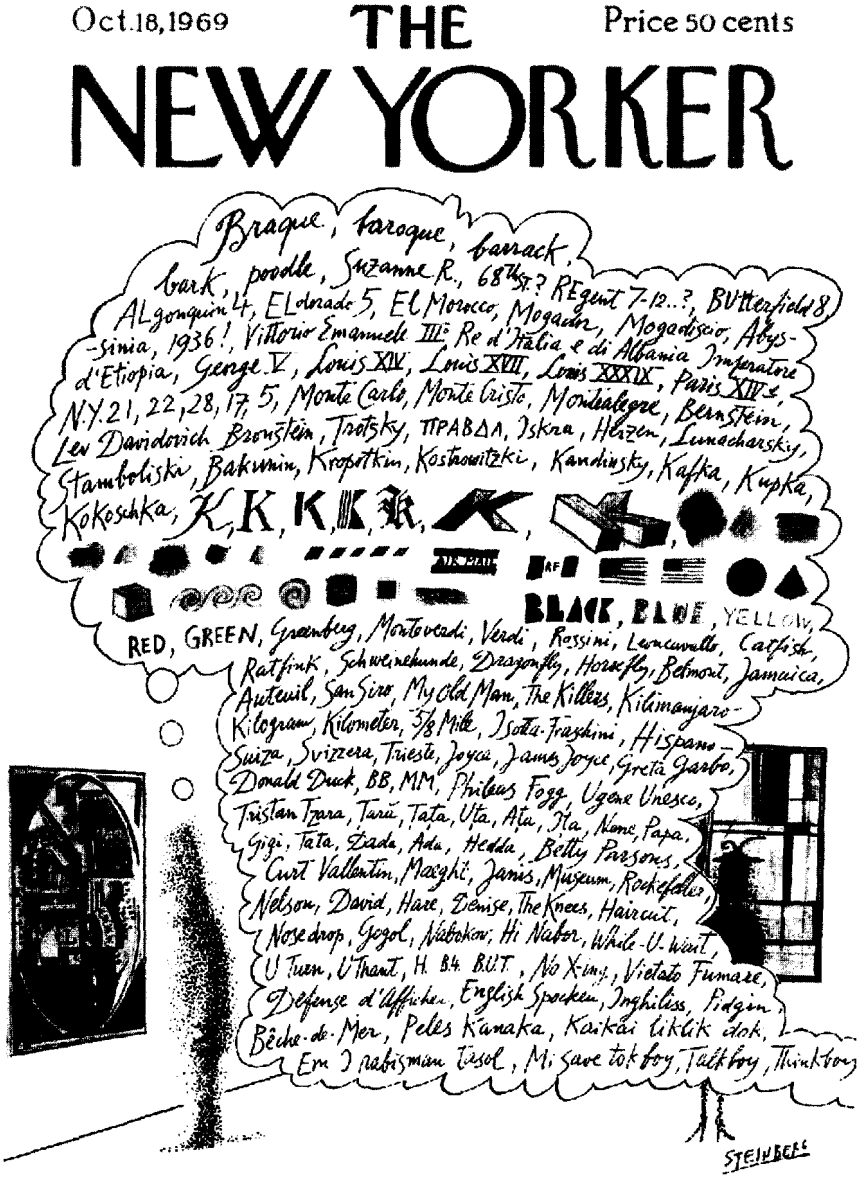
"The Zombic Hunch: Extinction of an Intuition?" in
PHILOSOPHY AT THE NEW MILLENNIUM, ed. Anthony O'Hear,
Cambridge Univ. Press, 2001, Royal Institute of Philosophy Supplement:
48, pp27-43

The Zombic Hunch: Extinction of an Intuition?

DANIEL DENNETT

Here is a good way of looking at the problem of consciousness:



If Saul Steinberg's 1967 *New Yorker* cover is the metaphorical truth about consciousness, what is the literal truth? What is going on in the world (largely in this chap's brain, presumably) that makes it the case that this gorgeous metaphor is so apt?

1. The Naturalistic Turn

Our conception of this question at the end of the twentieth century is strikingly different from the ways we might have thought about the same issue at the beginning of the century, thanks very little to progress in philosophy and very much to progress in science. Steinberg's *pointillist* rendering of our conscious man gives us a fine hint about the major advances in outlook that promise—to many of us—to make all the difference. What we now know is that each of us is an assemblage of trillions of cells, of thousands of different sorts. Most of the cells that compose your body are descendants of the egg and sperm cell whose union started you (there are also millions of hitchhikers from thousands of different lineages stowed away in your body), and, to put it vividly and bluntly, *not a single one of the cells that compose you knows who you are, or cares.*

The individual cells that compose you are alive, but we now understand life well enough to appreciate that each cell is a mindless mechanism, a largely autonomous micro-robot, no more conscious than a yeast cell. The bread dough rising in a bowl in the kitchen is teeming with life, but nothing in the bowl is sentient or aware—or if it is, then this is a remarkable fact for which, at this time, we have not the slightest evidence. For we now know that the 'miracles' of life—metabolism, growth, self-repair, self-defence, and, of course, reproduction—are all accomplished by dazzlingly intricate, but non-miraculous, means. No sentient supervisor is needed to keep metabolism going, no *élan vital* is needed to trigger self-repair, and the incessant nanofactories of replication churn out their duplicates without any help from ghostly yearnings or special life forces. A hundred kilos of yeast does not wonder about Braque, or about anything, but you do, and you are made of parts[1] that are fundamentally the same sort of thing as those yeast cells, only with different tasks to perform. Your trillion-robot team is gathered together in a breathtakingly efficient regime that has no dictator but manages to keep itself organized to repel outsiders, banish the weak, enforce iron rules of discipline—and serve as the headquarters of one conscious self, one mind. These communities of cells are fascistic in the extreme, but *your* interests and values

[1] Eukaryotic cells.

have almost nothing to do with the limited goals of the cells that compose you—fortunately. Some people are gentle and generous, others are ruthless; some are pornographers and others devote their lives to the service of God, and it has been tempting over the ages to imagine that these striking differences must be due to the special features of some *extra* thing—a soul—installed somehow in the bodily headquarters. Until fairly recently, this idea of a rather magical extra ingredient was the only candidate for an explanation of consciousness that even *seemed* to make sense. For many people, this idea (dualism) is *still* the only vision of consciousness that makes any sense to them, but there is now widespread agreement among scientists and philosophers that dualism is—must be—simply false: we are each *made of* mindless robots and nothing else, no non-physical, non-robotic ingredients at all.

But how could this possibly be? More than a quarter of a millennium ago, Leibniz posed the challenge to our imaginations with a vivid intuition pump, a monumentally misleading grandfather to all the Chinese Rooms (Searle), Chinese Nations (Block) and latter-day zombies.

> Moreover, it must be confessed that *perception* and that which depends upon it are *inexplicable on mechanical grounds*, that is to say, by means of figures and motions. And supposing there were a machine, so constructed as to think, feel, and have perception, it might be conceived as increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception. Thus it is in a simple substance, and not in a compound or in a machine, that perception must be sought for. (Leibniz, *Monadology*, 1714:parag. 17 [Latta translation]).

There is a striking *non sequitur* in this famous passage, which finds many echoes in today's controversies. Is Leibniz's claim epistemological—we'll *never understand* the machinery of consciousness—or metaphysical—consciousness *couldn't be* a matter of 'machinery'? His preamble and conclusion make it plain that he took himself to be demonstrating a metaphysical truth, but the only grounds he offers would—at best—support the more modest epistemological reading.[2] Somebody *might* have used Leibniz's wonderful

[2] Leibniz makes this particularly clear in another passage quoted in Latta's translation: 'If in that which is organic there is nothing but mechanism, that is, bare matter, having differences of place, magnitude and figure; nothing can be deduced or explained from it, except mechanism, that

Gulliverian image to *illustrate and render plausible*[3] the claim that although consciousness is—must be, in the end—a product of some gigantically complex mechanical system, it will surely be utterly beyond anybody's intellectual powers to explain how this is so. But Leibniz clearly intends us to treat his example as demonstrating the absurdity of the very idea that consciousness could be such an emergent effect of a hugely complex machine ('Thus it is in a simple substance, and not in a compound or in a machine, that perception must be sought for.') The same mismatch between means and ends haunts us today: Noam Chomsky, Thomas Nagel and Colin McGinn (among others) have all surmised, or speculated, or claimed, that consciousness is beyond all human understanding, a mystery not a puzzle, to use Chomsky's proposed distinction.[4] According to this line of thought, we lack the wherewithal—the brain power, the perspective, the intelligence—to grasp *how* the 'parts which work one upon another' could constitute consciousness. Like Leibniz, however, these thinkers have also hinted that they themselves understand the mystery of consciousness a little bit—just well enough to able to conclude that it *couldn't* be solved by any mechanistic account. And just like Leibniz, they have offered nothing, really, in the way of arguments for their pessimistic conclusions beyond a compelling image. When they contemplate the prospect they simply draw a blank. and thereupon decide that no further enlightenment lies down that path *or could possibly* lie down that path.

[3] It would not, of course, *prove* anything at all. It is just an intuition pump.

[4] Most recently, in the following works: Noam Chomsky, 'Naturalism and Dualism in the Study of Mind and Language', *Int. J. of Phil. Studies*, vol. 2, pp. 181–209 (his Agnes Cuming lecture of 1993), 1994. Thomas Nagel, 'Conceiving the Impossible and the Mind-Body Problem', *Philosophy*, **73,** 1998, pp. 337–52. Colin McGinn, *The Mysterious Flame: Conscious Minds in a Material World* (New York: Basic Books, 1999).

---

is, except such differences as I have mentioned. For from anything taken by itself nothing can be deduced and explained, except differences of the attributes which constitute it. Hence we may readily conclude that in no mill or clock as such is there to be found any principle which perceives what takes place in it; and it matters not whether the things contained in the 'machine' are solid or fluid or made up of both. Further we know that there ence of magnitude. Whence it follows that, if it is inconceivable how perception arises in any coarse 'machine', whether it be made up of fluids or solids, it is equally inconceivable how perception can arise from a fine 'machine'; for it our senses were finer, it would be the same as if we were perceiving a coarse 'machine', as we do at present.' [from *Commentatio de Anima Brutorum*, 1710, quoted in footnote in Latta, p. 228.]

Might it be, however, that Leibniz, lost in his giant mill, just couldn't see the woods for the trees? Might there not be a birds-eye view—*not* the first-person perspective of the subject in question, but a higher-level third-person perspective—from which, if one squinted just right, one could bring into focus the recognizable patterns of consciousness in action? Might it be that somehow the organization of all the parts which work one upon another yields consciousness as an emergent product? And if so, why couldn't we hope to understand it, once we had developed the right concepts? This is the avenue that has been enthusiastically and fruitfully explored during the last quarter century under the twin banners of cognitive science and functionalism—the extrapolation of *mechanistic naturalism* from the body to the mind. After all, we have now achieved excellent mechanistic explanations of metabolism, growth, self-repair, and reproduction, which not so long ago also looked too marvellous for words. Consciousness, on this optimistic view, is indeed a wonderful thing, but not *that* wonderful—not too wonderful to be explained using the same concepts and perspectives that have worked elsewhere in biology. Consciousness, from this perspective, is a relatively recent fruit of the evolutionary algorithms that have given the planet such phenomena as immune systems, flight, and sight. In the first half of the century, many scientists and philosophers might have agreed with Leibniz about the mind, simply because the mind seemed to consist of phenomena *utterly unlike* the phenomena in the rest of biology. The inner lives of mindless plants and simple organisms (and our bodies below the neck) might yield without residue to normal biological science, but nothing remotely mindlike could be accounted for in such mechanical terms. Or so it must have seemed until something came along in midcentury to break the spell of Leibniz's intuition pump; computers. Computers are mindlike in ways that no earlier artifacts were: they can control processes that perform tasks that call for discrimination, inference, memory, judgment, anticipation; they are generators of new knowledge, finders of patterns—in poetry, astronomy, and mathematics, for instance—that heretofore only human beings could even hope to find. We now have real world artifacts that dwarf Leibniz's giant mill both in speed and intricacy. And we have come to appreciate that what is well nigh invisible at the level of the meshing of billions of gears may nevertheless be readily comprehensible at higher levels of analysis—at any of many nested 'software' levels, where the patterns of patterns of patterns of organization (of organization of organization) can render salient *and explain* the marvellous competences of the mill. The sheer existence of computers has provided an existence proof of undeniable influence:

there are mechanisms—brute unmysterious mechanisms operating according to routinely well-understood physical principles—that have many of the competences heretofore assigned only to minds.

One thing we know to a moral certainty about computers is that there is nothing up their sleeves: no ESP or morphic resonance between the disk drives, no action-at-a-distance accomplished via strange new forces. The *explanations* of whatever talents computers exhibit are models of transparency, which is one of the most attractive features of cognitive science: we can be quite sure that *if* a computational model of *any* mental phenomena is achieved, it will inherit this transparency of explanation from its simpler ancestors.

In addition to the computers themselves, wonderful exemplars and research tools that they are, we have the wealth of new concepts computer science has defined and made familiar. We have learned how to think fluently and reliably about the cumulative effects of intricate cascades of micro-mechanisms, trillions upon trillions of events of billions of types, interacting on dozens of levels. Can we harness these new powers of disciplined imagination to the task of climbing out of Leibniz's mill? The hope that we can is, for many of us, compelling—even inspiring. We are quite certain that a naturalistic, mechanistic explanation of consciousness is not just possible; it is fast becoming actual. It will just take a lot of hard work of the sort that has been going on in biology all century, and in cognitive science for the last half century.

2. The Reactionaries

But in the last decade of the century a loose federation of reactionaries has sprung up among philosophers in opposition to this evolutionary, mechanistic naturalism. As already noted, there are the *mysterians*, Owen Flanagan's useful term for those who not only find this optimism ill-founded but also think that defeat is certain. Then there are those who are not sure the problem is insoluble, but do think they can titrate the subtasks into the 'easy problems' and the 'Hard Problem' (David Chalmers) or who find what they declare to be an Explanatory Gap (Joseph Levine) that has so far—and perhaps always will—defy those who would engulf the mind in one unifying explanation.[5] A curious anachronism found in many

[5] David Chalmers, 'Facing Up to the Problems of Consciousness', *J. Consc. Studies*, **2**, pp. 200–19, and *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, 1996). Joseph Levine, 'Materialism and Qualia: The Explanatory Gap', *Pacific Philosophical Quarterly*, **64**, pp. 354–61, 1983.

but not all of these reactionaries is that to the extent that they hold out any hope at all of solution to the problem (or problems) of consciousness, they speculate that it will come not from biology or cognitive science, but from—of all things—physics!

One of the first to take up this courtship with physics was David Chalmers, who suggested that a theory of consciousness should 'take experience itself as a fundamental feature of the world, alongside mass, charge, and space-time.' As he correctly noted, 'No attempt is made [by physicists] to explain these features in terms of anything simpler'[6] a theme echoed by Thomas Nagel:

> Consciousness should be recognized as a conceptually irreducible aspect of reality that is necessarily connected with other equally irreducible aspects—as electromagnetic fields are irreducible to but necessarily connected with the behaviour of charged particles and gravitational fields with the behaviour of masses, and vice versa.[7]

And Noam Chomsky:

> The natural conclusion ... is that human thought and action are properties of organized matter, like 'powers of attraction and repulsion', electrical charge, and so on.[8]

And Galen Strawson, who says, in a review of Colin McGinn's most recent book: 'we find consciousness mysterious only because we have a bad picture of matter' and adds:

> We have a lot of mathematical equations describing the behaviour of matter, but we don't really know anything more about its intrinsic nature, The only other clue that we have about its intrinsic nature, in fact. is that when you arrange it in the way that it is arranged in things like brains, you get consciousness.[9]

Not just philosophers and linguists have found this an attractive idea. Many physicists have themselves jumped on the bandwagon, following the lead of Roger Penrose, whose speculations about quantum fluctuations in the microtubules of neurons have attracted considerable attention and enthusiasm in spite of a host of

[6] Chalmers, 'Facing Up to the Problems of Consciousness', *J. Consc. Studies*, 2, pp. 200–19.

[7] Nagel, op.cit., p. 338.

[8] Chomsky, op. cit., 189. Chomsky is talking about the conclusion drawn by La Mettrie and Priestley, but his subsequent discussion, footnoting Roger Penrose and John Archibald Wheeler, makes it clear that he thinks this is a natural conclusion today, not just in early post-Newtonian days.

[9] Galen Strawson, 'Little Gray Cells,' *New York Times Book Review,* 7/11/99, p. 13.

problems.[10] What all these views have in common is the idea that some revolutionary principle of physics could be a *rival* to the idea that consciousness is going to be explained in terms of 'parts which work one upon another,' as in Leibniz's mill.

Suppose they are right. Suppose the Hard Problem—whatever it is—can only be solved by confirming some marvellous new and irreducible property of the *physics* of the cells that make up a brain. One problem with this is that the physics of your brain cells is, so far as we know, the same as the physics of those yeast cells undergoing population explosion in the dish. The differences in functionality between neurons and yeast cells are explained in terms of differences of cell anatomy or cytoarchitecture, not physics. Could it be, perhaps, that those differences in anatomy permit neurons to respond to physical differences to which yeast cells are oblivious? Here we must tread carefully, for if we don't watch out, we will simply reintroduce Leibniz's baffling mill at a more microscopic level—watching the quantum fluctuations in the microtubules of a single cell and not being able to see how any amount of *those* 'parts which work one upon another' could explain consciousness. If you want to avoid the bafflement of Leibniz's mill, the idea had better be, instead, that consciousness is an irreducible property that inheres, somehow, 'in a simple substance,' as Leibniz put it, 'and not in a compound or in a machine.' So let us suppose that, thanks to their physics, neurons enjoy a tiny smidgen (a quantum, perhaps!) of consciousness. We will then have solved the problem of how large ensembles of such cells—such as you and I—are conscious: we are conscious because our brains are made of the right sort of stuff, stuff with the micro-*je-ne-sais-quoi* that is needed for consciousness. But even if we had solved *that* problem, we would still have the problem illustrated by my opening illustration: how can cells, even *conscious* cells, that themselves know nothing about

[10] Incurable optimist that I am, I find this recent invasion by physicists into the domains of cognitive neuroscience to be a cloud with a silver lining: for the first time in my professional life, an interloping discipline beats out philosophy for the prize for combining arrogance with ignorance about the field being invaded. Neuroscientists and psychologists who used to stare glassy-eyed and uncomprehending at philosophers arguing about the fine points of *supervenience* and *intensionality-with-an-s* now have to contend in a similar spirit with the arcana of *quantum entanglement* and *Bose-Einstein condensates*. It is tempting to suppose that as it has become harder and harder to make progress in physics, some physicists have sought greener pastures where they can speculate with even less fear of experimental recalcitrance or clear contradiction.

art or dogs or mountains compose themselves into a thing that has conscious thoughts about Braque or poodles or Kilimanjaro? How can the whole ensemble be so knowledgeable of the passing show, so in touch with distal art objects (to say nothing of absent artists and mountains) when all of its parts, however conscious or sentient they are, are myopic and solipsistic in the extreme? We might call this the *topic*-of-consciousness question.

I suspect that this turn to physics looks attractive to some people mainly because they have not yet confronted the need to answer *this* question, for once they do attempt it, they find that a 'theory' that postulates some fundamental and irreducible sentience-field or the like has no resources at all to deal with it. Only a theory that proceeds in terms of how the parts work together in larger ensembles has any hope of shedding light on the topic question, and once theory has ascended to such a high level, it is not at all clear what use the lower-level physical sophistications would be. Moreover, there already are many models of systems that uncontroversially answer *versions* of the topic question, and they are all computational. How can the little box on your desk, whose parts know nothing at all about chess, beat you at chess with such stunning reliability? How can the little box driving the pistons attached to the rudder do a better job of steering a straight course than any old salt with decades at sea behind him? Leibniz would have been ravished with admiration by these mechanisms, which would have shaken his confidence—I daresay—in the claim that no mechanistic explanation of 'perception' was possible.[11]

David Chalmers, identifier of the Hard Problem, would agree with me, I think. He would classify the topic question as one of the 'easy problems'—one of the problems that *does* find its solution in terms of computational models of control mechanisms. It follows from what he calls the principle of organizational invariance.[12] Consider once again our *pointillist* gentleman and ask if we can tell from the picture whether he's a genuinely conscious being or a zombie—a philosopher's zombie that is behaviourally indistinguishable from a normal human being but is utterly lacking in consciousness. Even the zombie version of this chap would have a head full of

[11] A classic example of the topic problem in nature, and its ultimately computational solution, is Douglas Hofstadter's famous 'Prelude ... Ant Fugue' in *Gödel Escher Bach* (1979), the dialogue comparing an ant colony ('Aunt Hillary') to a brain, whose parts are equally clueless contributors to systemic knowledge of the whole. In his reflections following the reprinting of this essay in Hofstadter and Dennett, (eds), *The Mind's I*, (1981), he asks 'Is the soul more than the hum of its parts?'

[12] Chalmers, 1996, op. cit., esp. chapter 7.

dynamically interacting data-structures, with links of association bringing their sequels on-line, suggesting new calls to memory, composing on the fly new structures with new meanings and powers. Why? Because only a being with such a system of internal operations and activities could non-miraculously maintain the complex set of behaviours this man would no doubt exhibit, if we put him to various tests. If you want a theory of all that information-processing activity, it will have to be a computational theory, whether or not the man is conscious. According to Chalmers, where normal people have a stream of consciousness, zombies have a stream of unconsciousness, and he has argued persuasively that whatever explained the *purely informational competence* of one (which includes every transition, every construction, every association depicted in this thought balloon) would explain the same competence in the other. Since the literal truth about the mechanisms responsible for all the sworls and eddies in the stream, as well as the informational contents of the items passing by, is—*ex hypothesi*—utterly unaffected by whether or not the stream is conscious or unconscious, Steinberg's cartoon, a brilliant metaphorical rendering of consciousness, is exactly as good a metaphorical rendering of what is going on inside a zombie. (See, e.g., the discussion of zombie beliefs in Chalmers, 1996, pp. 203–5.)

3. An Embarrassment of Zombies

Must we talk about zombies? Apparently we must. There is a powerful and ubiquitous intuition that computational, mechanistic models of consciousness, of the sort we naturalists favour, *must leave something out*—something important. Just what must they leave out? The critics have found that it's hard to say, exactly: qualia, feelings, emotions, the what-it's-likeness (Nagel)[13] or the ontological subjectivity (Searle)[14] of consciousness. Each of these attempts to characterize the phantom residue has met with serious objections and been abandoned by many who nevertheless want to cling to the intuition, so there has been a gradual process of distillation, leaving just about all the reactionaries, for all their disagreements among themselves, united in the conviction *that there is a real difference between a conscious person and a perfect zombie*—let's call that intuition the *Zombic Hunch*—leading them to the thesis of

[13] Thomas Nagel, 1974, 'What is it Like to be a Bat?' *Phil. Review*, **83**, pp. 435–50.

[14] John Searle, *The Rediscovery of the Mind*, (MIT Press, 1992).

Zombism: that *the fundamental flaw in any mechanistic theory of consciousness is that it cannot account for this important difference.*[15] A hundred years from now, I expect this claim will be scarcely credible, but let the record show that in 1999. John Searle, David Chalmers, Colin McGinn, Joseph Levine and many other philosophers of mind don't just *feel the tug* of the Zombic Hunch (I can feel the tug as well as anybody), they *credit* it. They are, however reluctantly, Zombists, who maintain that the zombie challenge is a serious criticism. It is not that they don't recognize the awkwardness of their position. The threadbare stereotype of philosophers passionately arguing about how many angels can dance on the head of a pin is not much improved when the topic is updated to whether zombies—admitted by all to be imaginary beings—are (1) metaphysically impossible, (2) logically impossible, (3) physically impossible, or just (4) extremely unlikely to exist. The reactionaries have acknowledged that many who take zombies seriously have simply failed to imagine the prospect correctly. For instance, if you were surprised by my claim that the Steinberg cartoon would be an equally apt metaphorical depiction of the goings on in a zombie's head, you had not heretofore understood what a zombie is (and isn't). More pointedly, if you *still* think that Chalmers and I are just wrong about this, you are simply operating with a mistaken concept of zombies, one that is irrelevant to the philosophical discussion. (I mention this because I have found that many onlookers, scientists in particular, have a hard time believing that philosophers can be taking such a preposterous idea as zombies seriously, so they generously replace it with some idea that one *can* take seriously—but one that does not do the requisite philosophical work. Just remember, by definition, a zombie behaves *indistinguishably* from a conscious being—in all possible tests, including not only answers to questions [as in the Turing test] but psychophysical tests, neurophysiological tests—all tests that any 'third-person' science can devise.)

Thomas Nagel is one reactionary who has recoiled somewhat from zombies. In his recent address to this body, Nagel is particularly circumspect in his embrace. On the one hand, he declares that naturalism has so far failed us:

> We do not at present possess the conceptual equipment to understand how subjective and physical features could both be essential aspects of a single entity or process.

[15] In the words of one of their most vehement spokespersons, 'It all comes down to zombies.' Selmer Bringsjord, 'Dennett versus Searle: It All Comes Down to Zombies and Dennett is Wrong,' (APA December, 1994).

Why not? Because 'we still have to deal with the apparent conceivability of ... a zombie.' Notice that Nagel speaks of the *apparent* conceivability of a zombie. I have long claimed that this conceivability is *only* apparent; some misguided philosophers *think* they can conceive of a zombie, but they are badly mistaken.[16] Nagel, for one, agrees:

> the powerful intuition that it is conceivable that an intact and normally functioning physical human organism could be a completely unconscious zombie is an illusion.[17]

David Chalmers is another who is particularly acute in his criticisms of the standard mis-imaginations that are often thought to support the zombie challenge (his 1996 chapter 7, 'Absent Qualia, Fading Qualia, Dancing Qualia,' bristles with arguments against various forlorn attempts), but in the end, he declares that although zombies are in every realistic sense impossible, we 'non-reductive functionalists' still leave something out—or rather, we leave a job undone. We cannot provide '*fundamental* laws' from which one can deduce that zombies are impossible (p. 276 and elsewhere). Chalmers' demand for fundamental laws lacks the independence he needs if he is to support his crediting of the Zombie Hunch, for it arises from that very intuition: *if* you believe that consciousness sunders the universe in twain, into those things that have it and those that don't, *and* you believe this is a fundamental metaphysical distinction, then the demand for fundamental laws that enforce and explain the sundering makes some sense, but we naturalists think that this elevation of consciousness is itself suspect; supported by tradition and nothing else. Note that nobody these days would clamour for fundamental laws of the theory of kangaroos, showing why pseudo-kangaroos are physically, logically, metaphysically impossible. Kangaroos are wonderful, but not *that* wonderful. We naturalists think that consciousness, like locomotion or predation, is something that comes in different varieties, with some shared functional properties, but many differences, due to different evolutionary histories and circumstances. We have no use for fundamental laws in making these distinctions.

We are all susceptible to the Zombic Hunch, but if we are to credit

[16] Daniel Dennett, 1991, *Consciousness Explained*, New York and Boston: Little Brown, esp chapters 10–12; 1994, 'Get Real,' reply to 14 essays, in *Philosophical Topics*, **22**, no. I & 2, 1994, pp. 505–68; 1995, 'The Unimagined Preposterousness of Zombies,' *J. Consc. Studies*, **2**, pp. 322–36.

[17] Nagel, 1998, op. cit., p. 342.

it, we need a good argument, since the case has been made that it is a persistent cognitive illusion and nothing more. I have found no good arguments, and plenty of bad ones. So why, then, do so many philosophers persist in their allegiance to an intuition that they themselves have come to see is of suspect provenance? Partly, I think, this is the effect of some serious misdirection that has bedevilled communication in cognitive science in recent years.

4. Broad Functionalism and Minimalism

Functionalism is the idea that handsome is as handsome does, that 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 favour 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 noncommittal—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 fibres, 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], 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. To see what is at stake here, compare the neuromodulators to the food that is ingested by people. Psychologists and neuroscientists do not, as a rule, carefully inventory the food intake of their subjects, on the entirely plausible grounds that a serving of vanilla ice cream makes roughly the same contribution to how the brain goes about its tasks as a serving of strawberry ice cream. So long as there isn't any marijuana in the brownies, we can ignore the specifics of the food, and just treat it as a reliable energy source, the brain's power supply. This *could* turn out to be mistaken. It might turn out that psychologically important, if subtle, differences, hinged on whether one's subjects had recently had vanilla ice cream. Those who thought it did make a difference would have a significant empirical disagreement with those who thought it didn't, but this would not be disagreement between functionalists and anti-functionalists. It would be a disagreement between those who thought that functionalism had to be expanded downward to include the chemistry of food and those who thought that functionalism could finesse that complication. Consider the following:

> there may be various general neurochemical dispositions [based on the neuropeptide systems] that guide the patterning of thoughts that no amount of computational work can clarify. (Panskepp, 1998) Panskepp, J. 1998, *Affective Neuroscience: The Foundations of Human and Animal Emotions*, Oxford and NY, CUP.

This perfectly captures a widespread (and passionately endorsed) attitude, but note that there is nothing oxymoronic about a computational theory of neuromodulator diffusion and its effects, for instance, and pioneering work in 'virtual neuromodulators' and 'diffusion models of computational control' is well underway. Minds will turn out not to be simple computers, and their computational resources will be seen to reach down into the sub-cellular molecular resources available only to organic brains, but the theories that emerge will still be functionalist in the broad sense.

So within functionalism broadly conceived a variety of important controversies have been usefully playing themselves out, but an intermittently amusing side effect has been that many neuroscientists and psychologists who are rabidly anti-computer and anti-AI for various ideological reasons have mistakenly thought that philosophers' *qualia* and *zombies* and *inverted spectra* were useful weapons in their battles. So unquestioning have they been in their allegiance to the broad, bland functionalism of normal science, however, that they simply

haven't imagined that philosophers were saying what those philosophers were actually saying. Some neuroscientists have befriended *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 favour of qualia, in favour 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.

5. The future of an illusion

I do not know how long this ubiquitous misunderstanding will persist, but I am still optimistic enough to suppose that some time in the new century people will look back on this era and marvel at the potency of the visceral resistance[18] to the obvious verdict about the Zombic Hunch: it is an illusion.

[18] It is visceral in the sense of being almost entirely a-rational, insensitive to argument or the lack thereof. Probably the first to comment explicitly on this strange lapse from reason among philosophers was Lycan, in a footnote at the end of his 1987 book, *Consciousness* (MIT Press) that deserves quoting in full:

> On a number of occasions when I have delivered bits of this book as talks or lectures, one or another member of the audience has kindly praised my argumentative adroitness, dialectical skill, etc., but added that cleverness—and my arguments themselves—are quite beside the point, a mere exercise and/or display. Nagel (1979 [Preface to *Mortal Questions* Cambridge University Press]) may perhaps be read more charitably, but not much more charitably:
>
> > I believe one should trust problems over solutions, intuition over arguments .... [Well, excuuuuuse me!—WGL] If arguments or

Will the Zombic Hunch itself go extinct? I expect not. It will not survive in its current, toxic form but will persist as a less virulent mutation, still psychologically powerful but stripped of authority. We've seen this happen before. It still *seems* as if the earth stands still and the sun and moon go around it, but we have learned that it is wise to disregard this potent appearance as mere appearance. It still *seems* as if there's a difference between a thing at absolute rest and a thing that is merely not accelerating within an inertial frame, but we have learned not to trust this feeling. I anticipate a day when philosophers and scientists and laypeople will chuckle over the fossil traces of our earlier bafflement about consciousness: 'It still *seems* as if these mechanistic theories of consciousness leave something out, but of course that's an illusion. They do. in fact, explain everything about consciousness that needs explanation.'

If you find my prediction incredible, you might reflect on whether your incredulity is based on anything more than your current susceptibility to the Zombic Hunch. If you are patient and open-minded, it will pass.

> systematic theoretical considerations lead to results that seem intuitively not to make sense ... then something is wrong with the argument and more work needs to be done. Often the problem has to be reformulated, because an adequate answer to the original formulation fails to make the *sense* of the problem disappear (pp. x–xi).

If by this Nagel means only that intuitions contrary to ostensibly sound argument need at least to be explained away, no one would disagree (but the clause 'something is wrong with the argument' discourages that interpretation). The task of explaining away 'qualia'-based intuitive objections to materialism is what in large part I have undertaken in this book. If I have failed, I would like to be *shown why* (or, of course, presented with some new anti-materialist argument). To engage in further muttering and posturing would be idle. (pp. 147–8)