ELSEVIER

Cognition 79 (2001) 221–237

COGNITION

www.elsevier.com/locate/cognit

Are we explaining consciousness yet?

Daniel Dennett

Center for Cognitive Studies, Tufts University, Medford, MA 02155, USA

Received 27 August 2000; accepted 27 September 2000

Abstract

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

Keywords: Consciousness; Fame; Explaining

1. Clawing our way towards consensus

As the Decade of the Brain (declared by President Bush in 1990) comes to a close, we are beginning to discern how the human brain achieves consciousness. Dehaene and Naccache (in this volume) see convergence coming from quite different quarters on a version of the global neuronal workspace model. There are still many differences of emphasis to negotiate, and, no doubt, some errors of detail to correct, but there is enough common ground to build on. I agree, and will attempt to re-articulate

E-mail address: ddennett@tufts.edu (D. Dennett).

0010-0277/01/$ - see front matter
PII: S0010-0277(00)00130-X

this emerging view in slightly different terms, emphasizing a few key points that are often resisted, in hopes of precipitating further consolidation. (On the eve of the Decade of the Brain, Baars (1988) had already described a 'gathering consensus' in much the same terms: consciousness, he said, is accomplished by a "distributed society of specialists that is equipped with a working memory, called a *global workspace*, whose contents can be broadcast to the system as a whole" (p. 42). If, as Jack and Shallice (this volume) point out, Baars' functional neuroanatomy has been superceded, this shows some of the progress we've made in the intervening years.)

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

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

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

(2) There is no standard term for an event in the brain that carries information or content on some topic (e.g. information about color at a retinal location, information about a phoneme heard, information about the familiarity or novelty of other information currently being carried, etc.). Whenever some specialist network or smaller structure makes a discrimination, fixes some element of content, 'an information' in their sense comes into existence. 'Signal', 'content-fixation' (Dennett, 1991), 'micro-taking' (Dennett & Kinsbourne, 1992), 'wordless narrative' (Damasio, 1999), and 'representation' (see Jack and Shallice in this volume) are among the near-synonyms in use.

(3) We should be careful not to take the term 'top-down' too literally. Since there is no single organizational summit to the brain, it means only that such attentional amplification is not just modulated 'bottom-up' by features internal to the processing stream in which it rides, but also by *sideways* influences, from competitive, cooperative, collateral activities whose emergent net result is what we may lump together

and call top-down influence. In an arena of opponent processes (as in a democracy) the 'top' is distributed, not localized. Nevertheless, among the various competitive processes, there are important bifurcations or thresholds that can lead to strikingly different sequels, and it is these differences that best account for our pre-theoretical intuitions about the difference between conscious and unconscious events in the mind. If we are careful, we can use 'top-down' as an innocent allusion, exploiting a vivid fossil trace of a discarded Cartesian theory to mark the real differences that that theory mis-described. (This will be elaborated in my discussion of Jack and Shallice (this volume) below.)

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

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

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

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

> I would like to speak briefly about some of the advantages of the pandemonium model as an actual model of conscious behaviour. In observing a brain, one should make a distinction between that aspect of the behaviour which is available consciously, and those behaviours, no doubt equally important, but which proceed unconsciously. If one conceives of the brain as a pandemonium – a collection of demons – perhaps what is going on within the demons can be regarded as the unconscious part of thought, and what the demons are publicly shouting for each other to hear, as the conscious part of thought. (McCarthy, 1959, p. 147)

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

> specializations… form the building blocks for higher level intelligence… At the time of their origin, these specializations are tightly wired into the functional system they were designed to serve and are thus inaccessible to other programs or systems of the brain. I suggest that in the course of evolution these programs become more *accessible* to other systems and, in the extreme, may rise to the level of consciousness and be applied over the full realm of behavior or mental function. (p. 246)

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

2. Competition for clout

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

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

What a theory of consciousness needs to explain is how some relatively few contents become elevated to this political power, with all the ensuing *aftermath*, while most others evaporate into oblivion after doing their modest deeds in the ongoing projects of the brain. Why is this the task of a theory of consciousness? Because that is what conscious events do. They hang around, monopolizing time 'in the limelight'. We cannot settle for putting it that way, however. There is no literal searchlight of attention, so we need to explain away this seductive metaphor by explaining the functional powers of attention-*grabbing* without presupposing a single attention-*giving* source. This means we need to address two questions. Not just (1) How is this fame in the brain achieved? but also (2) – which I have called the Hard Question – And Then What Happens? (Dennett, 1991, p. 255). One may postulate activity in one neural structure or another as the necessary and sufficient condition for consciousness, but one must then take on the burden of the explaining why *that* activity ensures the political power of the events it involves – and this means taking a good hard look at how the relevant differences in competence might be enabled by changes in status in the brain.

Hurley (1998) makes a persuasive case for taking the Hard Question seriously in somewhat different terms: the Self (and its surrogates, the Cartesian *res cogitans*, the Kantian transcendental ego, among others) is not to be located by subtraction, by peeling off the various layers of perceptual and motor 'interface' between Self and World. We must reject the traditional 'sandwich' in which the Self is isolated from

the outside world by layers of 'input' and 'output'. On the contrary, the Self is large, concrete, and visible in the world, not just 'distributed' in the brain but spread out into the world. Where we act and where we perceive is not funneled through a bottleneck, physical or metaphysical, in spite of the utility of such notions as '*point* of view'. As she notes, the very content of perception can change, while keeping input constant, by changes in output (p. 289).

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

response space' (Frith, 2000, discussed in Jack and Shallice in this volume). This downstream capacity of one information to change the competitive context for whatever informations succeed it is indeed a fame-like competence, a hugely heightened influence that not only retrospectively distinguishes it from its competitors at the time but also, just as importantly, contributes to the creation of a relatively long-lasting Executive, not a place in the brain but a sort of political coalition that can be seen to *be in control* over the subsequent competitions for some period of time. Such differences in aftermath can be striking, perhaps never more so than those recently demonstrated effects that show, as Dehaene and Naccache (this volume) note, "the impossibility for subjects [i.e. Executives] to strategically use the unconscious information", in such examples as Debner and Jacoby (1994) and Smith and Merikle (1999) (discussed in Merikle, Smilek, & Eastwood in this volume).

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

The same point needs to be appreciated about consciousness, for this is where theorists' imaginations are often led astray: it is a mistake to go looking for an *extra* will-of-the-wisp property of consciousness that might be enjoyed by some events in the brain in spite of their not enjoying the fruits of fame in the brain. Just such a quest is attempted by Block (this volume), who tries to isolate 'phenomenality' as something distinct from fame ('global accessibility') but still worthy of being called a variety of consciousness. "Phenomenality is experience", he announces, but what does this mean? He recognizes that in order to keep phenomenality distinct from

global accessibility, he needs to postulate, and find evidence for, what he calls "phenomenality without reflexivity" – experiences that you don't know you're having.

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

But what then is left of the claim that phenomenality is experience? What is *experiential* (as contrasted with what?) about a discrimination that is not globally accessible? As the convolutions of Block's odyssey reveal, there is always the simpler hypothesis to fend off: there is *potential* fame in the brain (analogous to the dispositional status of poor Jim, the novelist) and then there is fame in the brain, and these two categories suffice to handle the variety of phenomena we encounter. Fame in the brain is enough.

3. Is there also a Hard Problem?

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

A model that, on the contrary, undertakes from the outset to address the Hard Question, assumes the obligation of accounting for the Subject in terms of "a collective dynamic phenomenon that does not require any supervision", as Dehaene and Naccache (this volume) put it. This risks seeming to leave out the Subject, precisely because all the work the Subject would presumably have done, once it had enjoyed the show, has already been parceled out to various agencies in the brain, leaving the Subject with nothing to do. We haven't really solved the problem of consciousness until that Executive is itself broken down into subcomponents that are themselves *clearly* just unconscious underlaborers which themselves work

(compete, interfere, dawdle, …) without supervision. Contrary to appearances, then, those who work on answers to the Hard Question are not leaving consciousness *out*, they are explaining consciousness by leaving it *behind*. That is to say, the only way to explain consciousness is to move beyond consciousness, accounting for the effects consciousness has when it is achieved. It is hard to avoid the nagging feeling, however, that there must be something that such an approach leaves out, something that lies somehow in between the causes of consciousness and its effects.

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

> If, in short, there is a community of computers living in my head, there had also better be somebody who is in charge; and, by God, it had better be me. (Fodor, 1998, p. 207)

> Of course the problem here is with the claim that consciousness is 'identical' to physical brain states. The more Dennett et al. try to explain to me what they mean by this, the more convinced I become that what they really mean is that consciousness doesn't exist. (Wright, 2000, fn. 14, Ch. 21)

> *Daniel Dennett is the Devil*… There is no internal witness, no central recognizer of meaning, and no self other than an abstract 'Center of Narrative Gravity' which is itself nothing but a convenient fiction… For Dennett, it is not a case of the Emperor having no clothes. It is rather that the clothes have no Emperor. (Voorhees, 2000, pp. 55–56)

This is not just my problem; it confronts anybody attempting to construct and defend a properly naturalistic, materialistic theory of consciousness. Damasio (1999) is one who has attempted to solve this pedagogical (or perhaps diplomatic)

problem by appearing to split the difference, writing eloquently about the Self, proclaiming that he is taking the Subject very seriously, even *restoring* the Subject to its rightful place in the theory of consciousness – while quietly dismantling the Self, breaking it into 'proto-selves' and identifying these in functional, neuroanatomic terms as a network of brain-stem nuclei (see Parvizi and Damasio in this volume). This effort at winsome redescription, which I applaud, includes some artfully couched phrases that might easily be misread, however, as conceding too much to those who fear that the Subject is being overlooked. One passage in particular goes to the heart of current controversy. They disparage an earlier account that "…dates from a time in which the phenomena of consciousness were conceptualized in exclusively behavioral, third-person terms. Little consideration was given to the cognitive, first-person description of the phenomena, that is, to the experience of the subject who is conscious." Notice that they do *not* say that they are now adopting a first-person perspective; they say that they are now giving more consideration to the 'first-person *description*' that subjects give. In fact, they are strictly adhering to the canons and assumptions of what I have called *heterophenomenology*, which is specifically designed to be a *third-person* approach to consciousness (Dennett, 1991, Ch. 4, p. 98). How does one take subjectivity seriously from a third-person perspective? By taking the *reports* of subjects seriously as reports of their subjective experience. This practice does not limit us to the study of human subjectivity; as numerous authors have noted, non-verbal animals can be put into circumstances in which some of their behavior can be interpreted, as Weiskrantz (1998) has put it, as 'commentaries', and Kanwisher (this volume) points out that in Newsome's experiments, for instance, the monkey's behavior is "a reasonable proxy for such a report".

It has always been good practice for scientists to put themselves in their own experimental apparatus as informal subjects, to confirm their hunches about what it feels like, and to check for any overlooked or underestimated features of the circumstances that could interfere with their interpretations of their experiments. (Kanwisher in this volume gives a fine example of this, inviting the reader into the role of the subject in rapid serial visual display (RSVP), and noting from the inside, as it were, the strangeness of the forced choice task: you find yourself thinking that 'tiger' would be as good a word as any, etc.) But scientists have always recognized the need to confirm the insights they have gained from self-administered pilot studies by conducting properly controlled experiments with naive subjects. As long as this obligation is met, whatever insights one may garner from 'first-person' investigations fall happily into place in 'third-person' heterophenomenology. Purported discoveries that cannot meet this obligation may inspire, guide, motivate, illuminate one's scientific theory, but *they* are not data – the beliefs of subjects about them are the data. Thus, if some phenomenologist becomes convinced by her own (first-) personal experience, however encountered, transformed, reflected upon, of the existence of a feature of consciousness in need of explanation and accommodation within her theory, her conviction that this is so is itself a fine datum in need of explanation, by her or by others, but the truth of her conviction must not be presupposed by science. There is no such thing as first-person science, so if you want to have a *science* of consciousness, it will have to be a third-person science of

consciousness, and none the worse for it, as the many results discussed in this volume show.

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

> I also take it that first-person data can't be expressed wholly in terms of third-person data about brain processes *and the like*. [my italics]… That's to say, no purely third-person description of brain processes *and behavior* [my italics] will express precisely the data we want to explain, though it may play a central role in the explanation. So 'as data', the first-person data are irreducible to third-person data. (p. 8)

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

The third-person approach is not antithetical to, or eager to ignore, the subjective nuances of experience; it simply insists on anchoring those subjective nuances to *something* – anything, really – that can be detected and confirmed in replicable experiments. For instance, Merikle et al. (this volume), having adopted the position that "with subjective measures, awareness is assessed on the basis of the observer's self-reports", note that one of the assumptions of this approach is that "information

perceived with awareness enables a perceiver to act on the world and to produce effects on the world". As contrasted to what? As contrasted to a view, such as that of Chalmers (1996) and Searle (1992), that concludes that consciousness *might* have no such enabling role – since a 'zombie' might be able to do everything a conscious person does, passing every test, reporting every effect, without being conscious. One of the inescapable implications of heterophenomenology, or of any third-person approach to subjectivity, is that one must dismiss as a chimera the prospect of a philosopher's zombie, a being that is behaviorally, objectively indistinguishable from a conscious person but not conscious. (For a survey of this unfortunate topic, see *Journal of Consciousness Studies*, *2*, 1995, "Zombie Earth: a symposium", including short pieces by many authors.)

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

4. But what about 'qualia'?

As Dehaene and Naccache note,

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

It is this informational superabundance, also noted by Damasio (1999) (see especially p. 93), that has lured philosophers into a definitional trap. As one sets out to answer the Hard Question ('And then what happens?'), one can be sure that no

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

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

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

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

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

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

This has left the mistaken impression in some quarters that the underlying idea of

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

This correction of overoptimistic minimalism has nothing to do with philosophers' imagined *qualia*. Some neuroscientists have thus muddied the waters by befriending qualia, confident that this was a term for the sort of functionally characterizable complication that confounds oversimplified versions of computationalism. (Others have thought that when philosophers were comparing zombies with conscious people, they were noting the importance of emotional state, or neuromodulator imbalance.) I have spent more time than I would like explaining to various scientists that their controversies and the philosophers' controversies are not translations of each other as they had thought but false friends, mutually irrelevant to each other. The principle of charity continues to bedevil this issue, however, and many scientists generously persist in refusing to believe that philosophers can be making a fuss about such a narrow and fantastical division of opinion. Meanwhile, some philosophers have misappropriated those same controversies within cognitive science to support their claim that the tide is turning against functionalism, in favor of qualia, in favor of the irreducibility of the 'first-person point of view' and so forth. This widespread conviction is an artifact of interdisciplinary miscommunication and nothing else. A particularly vivid exposure of the miscommunication can be found in the critics' discussion of Humphrey (2000). In his rejoinder Humphrey says

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

5. Conclusion

A neuroscientific theory of consciousness must be a theory of the Subject of consciousness, one that analyzes this imagined central Executive into component parts, none of which can itself be a proper Subject. The apparent properties of consciousness that only make sense as *features enjoyed by the Subject* must thus

also be decomposed and distributed, and this inevitably creates a pressure on the imagination of the theorist. No sooner do such properties get functionalistically analyzed into complex dispositional traits distributed in space and time in the brain, than their ghosts come knocking on the door, demanding entrance disguised as *qualia*, or *phenomenality* or *the imaginable difference between us and zombies*. One of the hardest tasks thus facing those who would explain consciousness is recognizing when some feature has *already* been explained (in sketch, in outline) and hence does not need to be explained again.

References

Brook, A. (in press). Judgments and drafts eight years later. In D. Ross, & A. Brook (Eds.), *Dennett's philosophy: a comprehensive assessment*. Cambridge, MA: MIT Press.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, 200–219.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Chalmers, D. (1999). First-person methods in the science of consciousness. *Consciousness Bulletin*, *Fall*, 8–11.

Crick, F. (1994). *The astonishing hypothesis: the scientific search for the soul*. New York: Scribner.

Damasio, A. (1999). *The feeling of what happens: body and emotion in the making of consciousness*, New York: Harcourt Brace.

Debner, J. A., & Jacoby, L.L. (1994). Unconscious perception: attention, awareness and control. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 304–317.

Dennett, D. (1988). Quining qualia. In A. Marcel, & E. Bisiach (Eds.), *Consciousness in modern science* (pp. 42–77). Oxford: Oxford University Press.

Dennett, D. (1991). *Consciousness explained*. Boston, MA: Little, Brown.

Dennett, D. (1994a). Real consciousness. In A. Revonsuo, & M. Kamppinen (Eds.), *Consciousness in philosophy and cognitive neuroscience* (pp. 55–63). Hillsdale, NJ: Lawrence Erlbaum.

Dennett, D. (1994b). Instead of qualia. In A. Revonsuo, & M. Kamppinen (Eds.), *Consciousness in philosophy and cognitive neuroscience* (pp. 129–139). Hillsdale, NJ: Lawrence Erlbaum.

Dennett, D. (1996a). Consciousness: more like fame than television [Bewusstsein hat mehr mit Ruhm als mit Fernsehen zu tun]. In C. Maar, E. Pöppel, & T. Christaller (Eds.), *Die Technik auf dem Weg zur Seele*Munich: Rowohlt.

Dennett, D. (1996b). Facing backwards on the problem of consciousness, commentary on Chalmers for *Journal of Consciousness Studies, 3* (1) (special issue, part 2), 4–6. Reprinted in J. Shear (Ed.), *Explaining consciousness – the 'Hard Problem'*. Cambridge, MA: MIT Press/Bradford Book, 1997.

Dennett, D. (1998a). The myth of double transduction. In S. Hameroff, A. W. Kaszniak, & A. C. Scott (Eds.), *International consciousness conference. Towards a science of consciousness II: the second Tucson discussions and debates* (pp. 97–107). Cambridge, MA: MIT Press.

Dennett, D., et al. (1998b). No bridge over the stream of consciousness, commentary on Pessoa et al. *Behavioral and Brain Sciences*, *21*, 753–754.

Dennett, D. (in press). The zombic hunch: the extinction of an illusion? *Philosophy* (special issue on philosophy at the Millennium).

Dennett, D., & Kinsbourne, M. (1992). Time and the observer: the where and when of consciousness in the brain. *Behavioral and Brain Sciences*, *15*, 183–247.

Fodor, J. (1998). Review of Steven Pinker's *How the mind works*, and Henry Plotkin's *Evolution in mind*. *London Review of Books*, Jan 22, 1998. Reprinted in Fodor, *In Critical Condition*. Cambridge, MA: MIT Press/Bradford Book, 1998.

Humphrey, N. (2000). How to solve the mind-body problem (with commentaries and a reply by the author). *Journal of Consciousness Studies*, *7*, 5–20.

Hurley, S. (1998). *Consciousness in action*. Cambridge, MA: Harvard University Press.

McCarthy, J. (1959). *Symposium on the mechanization of thought processes*. London: HMSO.
Penrose, R. (1989). *The emperors new mind: concerning computers, minds and the laws of physics*, Oxford: Oxford University Press.
Pessoa, L., Thompson, E., & Noë, A. (1998). Finding out about filling in: a guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, *21*, 723–802.
Petitot, J., Varela, F., Pachoud, B., & Roy, J.-M. (1999). *Naturalizing phenomenology: issues in contemporary phenomenology and cognitive science*. Stanford, CA: Stanford University Press.
Rozin, P. (1976). The evolution of intelligence and access to the cognitive unconscious. *Progress in psychobiology and physiological psychology* (Vol. 6, pp. 245–280). New York: Academic Press.
Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417–458.
Voorhees, B. (2000). Dennett and the deep blue sea. *Journal of Consciousness Studies*, *7*, 53–69.
Weiskrantz, L. (1998). Consciousness and commentaries. In S. R. Hameroff, A. W. Kaszniak, & A. C. Scott (Eds.), *Towards a science of consciousness II: the second Tucson discussions and debates* (pp. 11–25). Cambridge, MA: MIT Press.
Wright, R. (2000). *Nonzero: the logic of human destiny*. New York: Pantheon.