Notes on Daniel Dennett's *Consciousness Explained* (Boston: Little, Brown, 1991)

by Andrew Bailey, Philosophy Department, University of Guelph (abailey@uoguelph.ca)

1. Prelude: How Are Hallucinations Possible? 2
Part I: Problems and Methods 3
2. Explaining Consciousness 3
3. A Visit to the Phenomenological Garden 3
4. A Method for Phenomenology 4
A few questions for critical thought 5
Part II: An Empirical Theory of the Mind 6
5. Multiple Drafts versus the Cartesian Theater 6
6. Time and Experience 7
7. The Evolution of Consciousness 8
8. How Words Do Things With Us 10
9. The Architecture of the Human Mind 11
A few questions for critical thought 12
Part III: The Philosophical Problems of Consciousness 13
10. Show and Tell 13
11. Dismantling the Witness Protection Program 15
12. Qualia Disqualified 16
13. The Reality of Selves 18
14. Consciousness Imagined 19
A few questions for critical thought 20

1. Prelude: How Are Hallucinations Possible?

This section contains a 'theory of hallucinations' which is supposed to serve as a preamble to a general 'scientifically acceptable' theory of consciousness.

Because of *combinatorial explosion* "strong hallucinations are simply impossible" (with finite information-processing resources), according to Dennett. This raises the problem: how can we explain the convincing, multimodal hallucinations which do occur?

Powerful hallucinations/illusions are possible if the victim's "exploratory intentions and decisions" can be limited in advance (which constrains the combinatorial possibilities). Dennett (apparently) proposes that something like this effect could be achieved *automatically* in a neural system which operates by a kind of generate-and-test procedure. Such a system would produce hallucinations when "the hypothesis-generation side of the cycle … [operates] normally, while the data-driven side of the cycle … goes into a disordered or random or arbitrary round of confirmation and disconfirmation" (12). Dennett gives the analogy of the party game *Psychoanalysis* which, he says, produces "an illusion with no illusionist."

A similar model might also explain dreams, Dennett suggests.

Dennett concedes that this model is incomplete: we can explain hallucinations without an illusionist, but we still need to postulate a question-poser, and we need to explain the *consciousness* of dreams/hallucinations. Nevertheless he draws the following morals:

- Explanations of mental 'mysteries' can appeal to 'stupid' mechanisms.
- 'Engineering' constraints can provide useful clues.
- We do not represent to ourselves everything there is in the world/our environment—if we simply "assuage [all our] epistemic hunger" then it will *seem to us as if* our representation is a complete one.

PART I: PROBLEMS AND METHODS

2. Explaining Consciousness

Dennett suggests that the demystification of consciousness will be a good thing *even though* he admits that—on his view—'consciousness' is to a large degree a social construct, and so to topple that construct will be to eliminate aspects of consciousness as we currently understand it.

The mystery of consciousness, according to Dennett, is: "How can living physical bodies in the physical world produce such phenomena?" (25). Dennett breaks this down into the following problems:

- The medium of mental images—'mind stuff' (the 'stuff' of which the purple cow is made).
- The 'mind's eye'—the experiencer, or witness, of conscious mental events.
- The appreciation of (mental?) events—e.g. the enjoyment of the wine-taster; the *source of value*.
- The source of moral responsibility/agency.

In each case, the problem is that *the brain alone* seems unable to do/be these things.

But "dualism is forlorn," argues Dennett, and any theory of consciousness must be materialist. The standard problem with dualism is the problem of mind-body interaction—the 'Casper the Friendly Ghost' problem. "Accepting dualism is giving up" (37) because it is in principle incapable of being *explanatory*, claims Dennett—it does nothing to remove the mystery, but instead preserves it.

Dennett's methodological ground-rules:

- No wonder tissue allowed.
- No feigning anaesthesia.
- No nitpicking about empirical details.

3. A Visit to the Phenomenological Garden

A theory of consciousness will be a theory of phenomenology—of "all the items … that inhabit our conscious experience":

1) Experiences of the 'external' world (e.g. perceptions, kinaesthesia).
2) Experiences of the 'internal' world (e.g. dreams, ideas, thoughts).
3) Experiences of emotion/affect (e.g. bodily sensations, emotions, feelings).

But, according to Dennett, there are no "uncontroversial experts on the nature of the things that swim in the stream of consciousness" (45) … and that includes ourselves—we do not have the introspective authority we tend we think we have (Dennett argues).

- We can be surprised about the behaviour of our own senses (e.g. touching things through a wand, the lack of colour in peripheral vision).
- We can provide initial, partial analyses of apparently 'ineffable' sensations (e.g. the sound of a guitar string).
- We can discriminate perceptual information but have no idea *from the experience* how we do so (e.g. the gaps between words in speech). More generally, acquaintance with our own phenomenology is (often) totally unexplanatory (e.g. the nature of laughter).
- We can show that (at least some) phenomenology is irrelevant to cognition (e.g. the difference between listening to what is said and understanding it).

A good case study in this: Dennett's arguments that vision is not like pictures in the head (pp. 52–55).

But Dennett does not *deny* phenomenology (he says): "There must be something right about the idea of

mental imagery, and if 'pictures in the head' is the wrong way to think about it, we will have to find some better way to think about it" (58). Mental imagery can have powerful effects (e.g. hearing music by reading a score, evolutionary fitness of pain). He is—he claims—in search of "a materialist account that does justice to all these phenomena" (65).

4. A Method for Phenomenology

So we want to study the 'inhabitants' of consciousness, but our first-personal judgements about those entities are insufficiently rigorous (and certainly not 'incorrigible,' immediate,' etc.) … though Dennett does seem prepared to admit that we have 'some' privileged access.

Dennett diagnoses this failure of intersubjective phenomenal agreement as a result of the nature of *introspection*: introspection is not a matter of *observing* inner entities (on the model of perception), but is tacitly a kind of *theorizing* about our experience. E.g. we don't 'just see' that there are coloured patches throughout our visual field (since there aren't)—we *assume* that this must be so.

So 'pure' phenomenology does not work … because for Dennett there is no such thing. There is no such thing as neutrally observing the contents of our own consciousness. But we still need a method for neutrally describing the contents of consciousness to provide the data for our theory—Dennett proposes an 'impure' phenomenological method, which he calls *heterophenomenology*. The point of heterophenomenology is to provide us with a theory-neutral account of the data to be explained (or explained away) by an adequate scientific theory of consciousness.

NB: this method—as befits its scientific status—is a *third-person* (as opposed to first-person) methodology. It works by capturing *verbal behaviour* which purports to be reports of consciousness.

The method:

a) Record the sound-producing (or otherwise quasi-verbal—e.g. button-pushing) behaviour of the subject.
b) Transcribe this recording as a sequence of linguistic behaviours—i.e. *interpret* the behaviour as verbal. This, Dennett suggests, can be done in an 'objective' way (e.g. the three typists).
c) Interpret these linguistic behaviours as a sequence of *speech acts* by adopting the *intentional stance* towards the subject—i.e. treating them *as if* they are a rational agent with intentional (meaningful) states. According to Dennett, we preserve theory-neutrality here by treating this intentional interpretation of the subject's utterances as analogous to a work of fiction (i.e. as a self-consistent world which may not correspond to any existing reality).

"The subject's heterophenomenological world will be a stable intersubjectively confirmable theoretical posit" (81). The heterophenomenological description is neutral with respect to a) the mode in which the contents described are represented, and b) the truth or falsity of those contents (cp. Feenomanology).

These are the data for the science of consciousness. The theorist's task is then to discover the things in the real world which correspond most closely to those entities described in the heterophenomenological world—these things (whatever they are) are the real stuff of consciousness. Dennett says that we must find entities that are 'similar enough' to the heterophenomena, but that nevertheless the resemblance might still be 'minor'—we might be *mistaken* about many of the things we believe about our own consciousness. In particular, e.g., the purple cow might turn out to be brain events (85).

So, what are mental images? What kind of physical thing/process might correspond to—really be—the mental images we postulate in our own consciousness? Dennett gives the illustrative example of visual processing in Shakey the robot: its processes are suggestively close to human visual processing—it does some of what we do—but there are no 'pictures' in its head. Its processes are unmysterious, mechanical, 'stupid'—

to talk of 'images' in its head is merely metaphorical. This would be so *even if* Shakey was itself convinced that it had mental images, and could only stubbornly report that it introspected them.

So the heterophenomenological data are what need to be explained … but they may in fact be largely fictional ('confabulations'). This fictional status, however, does not take away from their scientific value, Dennett argues.

A few questions for critical thought

1) Does Dennett show conclusively that dualism is really "forlorn"?
2) Does Dennett make a good case that 'pure' phenomenology is impossible?
3) Does Dennett persuade you that introspection is theoretical and not (purely?) observational? Can he allow that it have *any* observational component?
4) Are there any data of consciousness which are simply theoretically non-negotiable: are there any introspectively supplied fixed points for the science of consciousness? What does/can Dennett say about this?
5) Can any third-person method replace first-personal observations? What implications does the answer to this question have?
6) How adequate is the method of heterophenomenology? Exactly what is its 'neutrality' supposed to consist in?
7) How similar is similar enough, in finding real world counterparts for heterophenomenological entities?
8) How plausible do you find Dennett's Shakey example? How, if at all, does it affect your intuitions about mental images?

PART II: AN EMPIRICAL THEORY OF THE MIND

5. Multiple Drafts versus the Cartesian Theater

People are observers—they have a point of view—but, Dennett argues, *there is no observer within the brain*. Points of view are simply not that fine-grained. This (apparently) is just because "there is no single point in the brain where all information funnels in" (102–103): but taken seriously, Dennett thinks, this simple fact has profound implications. In particular, he thinks that it means there is *no clear fact of the matter* about whether a particular stage of neural processing is conscious/experienced or not: there is no 'Continental Divide' in the brain that marks the difference between the pre-experiential and the post-experiential.

The myth that there is a point in the brain where 'consciousness happens' is what Dennett calls the Cartesian Theatre, and—combined with materialism—results in what he calls Cartesian materialism. Dennett is setting out to overthrow this myth and to replace it with a different model: the Multiple Drafts model.

A first take on multiple drafts: all mental activity is "accomplished in the brain by parallel, multitrack processes of interpretation and elaboration. … Information entering the nervous system is under continuous 'editorial revision'. … What we actually experience is a product of many processes of interpretation—editorial processes, in effect. … Feature detections or discriminations *only have to be made once*. That is, once a particular 'observation' of some feature has been made, the information content thus fixed does not have to be sent elsewhere to be *re*discriminated by some 'master' discriminator. … These spatially and temporally distributed content-fixations in the brain are precisely locatable in space and time, but their onsets do *not* mark the onset of consciousness of their content. … [I]t is a confusion, as we shall see, to ask *when it becomes conscious*. These distributed content-discriminations yield, over the course of time, something *rather like* a narrative stream or sequence, which can be though of as subject to continual editing by many processes distributed around in the brain. … This stream of contents is only rather like a narrative because of its multiplicity: at any one point in time there are multiple 'drafts' of narrative fragments at various stages of editing in various places in the brain. Probing this stream at different places and times produces different effects, precipitates different narratives from the subject. … Most important, the Multiple Drafts model avoids the tempting mistake of supposing that there must be a single narrative … that is canonical—that is the *actual* stream of consciousness of the subject." (111–113).

So:

a) Content-fixation is spatially and temporally distributed in the brain, and contents are never re-presented in a central 'Cartesian Theatre.'
b) Content-fixation is subject to an on-going editorial process of revision, and there is no particular (spatial or temporal) point in this process at which the contents *become conscious*—there simply is no fact of the matter about which 'are conscious' and which are not … according to Dennett this question does not even make sense.
c) This distributed process of content-fixation results in a set of *multiple drafts of narrative fragments* (narratives, presumably, about the organism's environment, for example), and this sequence of drafts is in on-going editorial flux, never resolving into a single canonical narrative.
d) 'Probing' the organism at a particular stage in this processing produces a particular behavioural response (including particular heterophenomenological reports), but this response cannot be thought of as a report or symptom of *the* stream of consciousness of the organism—a different probe at a different stage might have resulted in a quite different, but equally 'real,' report. "[T]here are no fixed facts about the stream of conscious independent of particular probes" (138).

A concrete example: the colour phi phenomenon, where red and green dots flashing in succession give the appearance of a moving dot changing colour—the puzzle is that the colour change appears to magically occur *before* the green dot has flashed. How do Cartesian Theatre and Multiple Drafts models account for

this?

Dennett introduces a distinction (which has subsequently generated a lot of discussion) between *Orwellian* and *Stalinesque* revisions.

- Orwellian revision is *post-experiential*: you first have a (veridical) experience, and then this is replaced by a fake memory—you come to believe that your experience was other than it really was.
- Stalinesque revision is *pre-experiential*: your experience is correctly remembered, but was itself falsidical—the editing of your perceptual data, for example, introduced falsehoods even before that perception became conscious.

Dennett (apparently) then argues roughly as follows:

i. If Cartesian materialism were true, there would have to be a fact of the matter with respect to whether the colour phi phenomenon were Orwellian or Stalinesque.
ii. But, experimentally (or even heterophenomenologically), there is no conceivable way of distinguishing between an Orwellian or Stalinesque mechanism for colour phi.
iii. "So, in spite of first appearances, there is really only a verbal difference between the two theories. … This is a difference that makes no difference" (125).
iv. [Hence the CT model is wrong and MD right?]

That is, there *is no real distinction* between Orwellian and Stalinesque revision in the brain—hence there is no real distinction between what you experience and what you remember having experienced. "[W]hat happened (in consciousness) is simply whatever you remember to have happened" (132).

Dennett then confronts the objection that this argument is too verificationist: that there can be no *experimental* confirmation of certain facts about consciousness (if that is so) does not mean that those facts do not exist (i.e. the facts required by CT models could be unverifiably present). But what could such facts be? Dennett seems to assume that they would be facts about 'phenomenological projection' into 'phenomenal space' for the benefit of some 'inner observer': he attacks this notion as ill-conceived. Among other things, he argues that there is no need to 'fill in' the missing 'frames' of the experienced motion—instead, the brain simply *judges that* motion and colour change has taken place. These judgements can *represent* a colour change from red to green without actually taking place *between* the red flash and the green flash—that is, "the representation of time in the brain does not always use time-in-the-brain" (131). Furthermore, he claims that the CT entails a "metaphysically dubious" notion of "the objectively subjective" (132)—a way things actually seem to you even if they don't seem to seem that way to you! For Dennett (subjective) seeming and (subjective) judging do not come apart—he calls this "first-person operationalism" (132).

6. Time and Experience

Dennett begins this chapter with more examples which are especially susceptible to explanation by the MD model: metacontrast and the cutaneous rabbit. What these kinds of cases have in common (and also in common with the colour phi case) is that later content-fixations can drive out earlier ones without the whole sequence having to be played again, and without forcing us to say definitively that the earlier contents were (or were not) conscious before being replaced.

This gives rise to a key issue, for Dennett: how the brain represents time. On the CT model, contents enter consciousness at particular times (like train cars passing through a clearing), and thus it is hard to see how the brain can represent time *except* in terms of the order in which these contents enter consciousness—that is, on the CT view, for A *to be represented as* experienced earlier than B is for it to *be* experienced earlier than B. (This is—arguably—what causes puzzles with colour phi and the cutaneous rabbit.) By contrast, on the MD model, the brain represents the temporal properties of events in the world in a distributed manner, and there is no need—or biological plausibility—for it to use time-in-the-brain to signal these temporal properties. Thus, the content A may be fixed later than the content B but still represent an earlier event than B. "In general, we

must distinguish features of represent*ings* from the features of represent*eds*" (147).

Dennett argues for this on 'engineering' grounds: roughly, signals from one and the same event enter the brain at different times, and are processed at different rates, and it would be highly inefficient for the brain to wait for the last signal to arrive and the last content to be fixed and then to 'synchronise' and 'reorder' them into the correct sequence and 'play them again.' The brain must use a large number of 'anticipatory strategies' to guide action, and there is no time to buffer perceptual information. "[T]here *must* be temporal smearing—on a small scale—*because* there must be spatial smearing (on a small scale) of the point of view of the observer" (152).

(Of course, as Dennett admits, there are constraints on the way time can be 'smeared' in the brain. First, perceptual timing may be what determines content: we see a moving dot against a black background only by seeing it first on the left and then later on the right. Only after the fact of this temporal ordering of perceptual data has been represented can *this* representation be used in 'temporally sloppy' ways. Second, representings may need to meet a particular deadline to be useful for guiding action, and so need to 'occur' within that temporal window.)

So how might the brain 'date stamp' its spatially and temporally distributed contents? Dennett suggests "content-sensitive settling," rather like the way a film soundtrack gets synchronized with its frames. Note that this 'date stamping' is not done to allow the contents to be *reordered* in some Cartesian Theatre: the temporal judgement *is* the experienced order of those contents, according to Dennett.

C.f. Libet's 'backwards referral in time' experiments: hand-tingles that start in the cortex would be expected to be felt quicker than those that are actually started in the hand (as the signals have to travel further), but in fact are felt *sooner*, suggesting that the tingle is "referred backwards in time" (155). This has been thought to show that materialism must be false, but Dennett argues that it does not—it just shows something suggestive, but inconclusive, about how the brain represents time (suggestive because it seems that it does so in different ways for different modalities—e.g. vision vs. fingertips).

C.f. Libet's experiments suggesting that "consciousness lags behind the brain processes that actually control your body" (163). Dennett suggests that this putative experimental data—about the 'subjective sequence' of events, compared with an 'objective sequence'—is incoherent, since there is no single 'time of occurrence' of internal representations. Something similar is, says Dennett, the problem with the case of 'Grey Walter's precognitive carousel' experiment.

Dennett considers, almost in passing, an important variant on the CT model: "On this model, an element of content becomes conscious at some time *t*, not by entering some functionally defined and anatomically located system, but by changing state right where it is: by acquiring some property or by having the intensity of one of its properties boosted above some threshold" (166). Dennett complains that this picture requires not only that conscious states have some special property but that the brain can discriminate this property. "We can put the crucial point as a question: What would make *this* sequence the stream of consciousness? There is no one inside, *looking at* the wide-screen show displayed all over the cortex, even if such a show is discernible by *outside* observers. What matters is the way those contents get utilized by or incorporated into the processes of the ongoing control of behavior, and this *must* be only indirectly constrained by cortical timing. What matters, once again, is not the temporal properties of the representings, but the temporal properties *represented*, something determined by how they are 'taken' by subsequent processes in the brain" (166).

7. The Evolution of Consciousness

Dennett turns to evolution to help him provide a more detailed positive account of the MD model: he sets

out to "think about the evolution of brain mechanisms for doing this and that, and see if anything that emerges gives us a plausible mechanism for explaining some of the puzzling 'behaviors' of our conscious brains" (172).

Dennett tells the following story:

i) First simple replicators appeared. With them came the first *reasons*: we "can nonarbitrarily assign them certain interests—generated by their defining 'interest' in self-replication" (173). This leads to *points of view*, from which "the world's events can be roughly partitioned into the favorable, the unfavorable, and the neutral" (174). And this leads to the importance of *boundaries*—the difference between 'inside' and 'outside'.

ii) Now these replicators need to be able to control their behaviour in accordance with their interests—they must develop *anticipation machines* to perform this guidance function … i.e., brains. At first, they are capable of only proximal anticipation of the immediate future, but they evolve to get more and more distal—this evolution takes place through better and better hardwiring (e.g. our visual symmetry-detection mechanism).

iii) One of these anticipation mechanisms is the *orienting response*, in which some alarm causes the organism to stop what it's doing for a moment and focus on scanning its environment for the presence of anything surprising. "These brief episodes of interruption and heightened vigilance are not [necessarily] themselves episodes of human-style 'conscious awareness' … but they are probably the necessary precursors, in evolution, of our conscious states" (180). Dennett speculates that these moments of vigilance proved so useful that organisms learned to initiate them themselves, and gradually a new strategy evolved: "the strategy of acquiring information 'for its own sake,' just in case it might prove valuable someday" (180–181). This Dennett calls "the birth of … epistemic hunger" (181), and suggests that the ventral brain evolved in mammals to accommodate it.

iv) So far this has been a story of genetic evolution. Now a second level of evolution comes into play: *postnatal design-fixing* (i.e. learning or development—phenotypic plasticity). According to Dennett, the only non-mysterious way in which this could occur is by "a process of evolution by natural selection that takes place within the individual" (183)—as such, it is a genetically evolved capacity for organisms to redesign their own brains. And this new capacity feeds back into genetic evolution and speeds it up via the Baldwin Effect (roughly, because it allows larger portions of a population to learn a Good Trick and thus creates new selection pressures making the Good Trick more salient for genetic evolution)—in this way, Good Tricks are moved into the genome.

v) What accounts for the vastly increased capacity for learning in human beings (compared to other animals)? Dennett speculates that it is connected to the rise of 'autostimulation.' There was probably a time, he says, in the evolution of language, "when vocalizations served the function of eliciting and sharing useful information" (195) within communities of hominids. Sometimes, Dennett suggests, a hominid might have mistakenly asked a question when no one was around, and surprised itself by answering its own question. This would lead to hominids asking themselves questions in order to 'broadcast' information from one part of the brain to other parts which lack the proper internal connections. This would quite quickly evolve—via the Baldwin Effect—into silent autostimulation. "This private talking-to-oneself behavior might well not be the best imaginable way of amending the existing functional architecture of one's brain, but it would be a close-to-hand, readily discovered enhancement, and that could be more than enough" (197). Something similar might have evolved from drawing pictures for oneself, Dennett suggests, leading to private diagrams 'in the mind's eye.'

vi) For human beings, there is also a third medium of evolution: memetic evolution. Memes are, roughly, ideas which replicate themselves in the competitive environment of culture, and hence face selective pressures. (Notice that what matters, on this model, is fitness *for the meme* rather than the degree of fitness which memes bestow on us, their carriers. As Dennett puts it, "A scholar is just a library's way of making another library" (202). On the whole, fitness for memes is their *attractiveness to us*, and thus their usefulness … but this is not always true (e.g. the conspiracy meme or the faith meme).) According to Dennett, memes parasitize the brain, and in fact "human minds are themselves to a very great degree the creations of memes. … Meme evolution has the potential to contribute

> remarkable design-enhancements to the underlying machinery of the brain—at great speed, compared to the slow pace of genetic R and D" (207–208).

Dennett, hence, defends the following hypothesis about consciousness:

> Human consciousness is *itself* a huge complex of memes (or more exactly, meme-effects in brains) that can best be understood as the operation of a "*von Neumannesque*" virtual machine *implemented* in the *parallel architecture* of a brain that was not designed for any such activities. The powers of this *virtual machine* vastly enhance the underlying powers of the organic *hardware* on which it runs, but at the same time many of its most curious features, and especially its limitations, can be explained as the *byproducts* of the *kludges* that make possible this curious but effective *reuse* of an existing organ for novel purposes. (210)

Dennett apparently feels particularly compelled to explain the 'Joycean' *seriality* of consciousness, since "the architecture of the brain, in contrast, is massively parallel, with millions of simultaneously active channels of operation" (214): this is why he introduces the notion of a von Neumannesque virtual machine. "We *know* there is something at least *remotely like* a von Neumann machine in the brain, because we know we have conscious minds 'by introspection' and the minds we thereby discover are at least this much like von Neumann machines: They were the inspiration for von Neumann machines!" (215). Consciousness, in other words, is a 'user illusion.' And the von Neumann machine is 'programmed' in individuals by memetic evolution.

A second key aspect of the virtual machine, apparently, is that it yokes " independently evolved specialist organs together in common cause…. It creates a *virtual captain* of the crew, without elevating any one of them to long-term dictatorial power. Who's in charge? First one coalition and then another, shifting in ways that are not chaotic thanks to good meta-habits that tend to entrain coherent, purposeful sequences rather than an interminable helter-skelter power grab" (228).

Dennett says it follows from all this that: "human consciousness (1) is too recent an innovation to be hard-wired into the innate machinery, (2) is largely a product of cultural evolution that gets imparted to brains in early training, and (3) its successful installation is determined by myriad microsettings in the plasticity of the brain, which means that its functionally important features are very likely to be invisible to neuroanatomical scrutiny in spite of the extreme salience of the effects" (219).

8. How Words Do Things With Us

This chapter is designed to "expose and neutralize another source of mystification: the illusion of the Central Meaner" (228). That is, 'who' 'decides' what we 'mean' to say? Dennett argues [apparently] that it does no good to postulate a 'Conceptualizer' module to explain this—the Conceptualizer (if it is to be explanatory at all) must 'issue instructions' to the Formulator module, and to do this it will have to compose the correct preverbal messages; this formulation of instructions will require its own explanation, leading to a vicious infinite regress of meaners.

Instead, Dennett proposes a model based on "a pandemonium of word demons" (237). Large numbers of parallel sentence-formation processes occur simultaneously, each one consisting of 'demons' modulating streams of random nonsense into various different sentences, and then these sentences 'compete' with each other to be uttered.

The key issue then is: "how is this tournament of words judged?" (238) How is the 'most suitable' utterance chosen, if not by a Central Meaner? Dennett suggests it is a process of 'constraint satisfaction' "that involves the collaboration, partly serial, partly in parallel, of various subsystems none of which is capable on its own of performing—or ordering—a speech act" (239). Dennett is—perhaps appropriately—agnostic about the

details of this process, but a central feature of this kind of model is that it is *non-bureaucratic*: it does not involve planning and instructions from above, but instead shows how meaning emerges from simpler, more 'stupid' processes.

A consequence of this kind of pandemonium model is that we do not have any kind of privileged access to our own meanings. "*Probably*, if I said it (and I heard myself say it, and I didn't hear myself rushing in with any amendments), I meant it, and it probably means what it seems to mean—to me" (246). Occasionally, we may consciously 'try out' a sentence before we speak it, but normally we learn what we are going to say—which sentence as won the competition—at the same time as our audience.

Dennett provides partial evidence for these kinds of views from speech pathologies.

He points out that the story he has told for speech acts goes for all varieties of intentional action. "We must build up … resistance to the temptation to explain *action* as arising from the imperatives of an internal action-orderer who does too much of the specification work. As usual, the way to discharge an intelligence that is too big for our theory is to replace it with an ultimately mechanical fabric of semi-independent semi-intelligences acting in concert. … Our actions generally satisfy us; we recognize that they are in the main coherent, and that they make appropriate, well-timed contributions to our projects as we understand them. So we safely assume them to be products of processes that are reliably sensitive to ends and means. That is, they are rational, in *one* sense of the word…. But that does not mean they are rational in a narrower sense: the product of serial reasoning." (251).

9. The Architecture of the Human Mind

Dennett begins this chapter with a thumbnail sketch of his theory so far (253–254). He then sets out to connect his theory with recent empirical work in neuroscience and cognitive psychology (and with some recent suggestions from philosophers). Along the way he entertains "the supposition that it is proving to be fiendishly difficult—but not impossible—to figure out how the brain works, in part because it was designed by a process that can thrive on multiple, superimposed functionality, something systematically difficult to discern from the perspective of reverse engineering" (273).

He then addresses the fundamental question: why is his model a model *of consciousness*? He answers this by describing the powers of the 'Joycean machine'. It (he says) deals with the "significant problems of self-control created by the proliferation of simultaneously active specialists" (277), and in turn gives us capacities which exceed those of the specialists (e.g. looking *for* something, rather than just looking *at* things), especially enhanced powers of anticipation and recollection. It 'broadcasts' information to the different specialists, "permitting *any* of the things one has learned to make a contribution to *any* current problem" (278), and thus [somehow] allows for the 'background' or 'context' required for sophisticated understanding.

Dennett concludes: "Anyone or anything that has such a virtual machine as its control system is conscious in the fullest sense, and is conscious *because* it has such a virtual machine" (281).

What about the zombie problem? Couldn't we conceive of a being with a Joycean machine but without consciousness? Dennett responds that he has succeeded in shifting the burden of proof onto those who think they can imagine such a thing. Part III of the book is (according to Dennett) designed to take on these challengers.

A few questions for critical thought

1) Does it follow from the fact that there is no single place in the brain where everything 'comes together' that there is no fact of the matter about whether a particular stage of neural processing is conscious or not? If this does not follow, then what other premises does Dennett need to make this case? Are these premises available to him?
2) Is there a difference between a) claiming that there is no Cartesian Theatre; b) claiming that there is no experiential Continental Divide in the brain; and c) claiming that there is no 'single narrative' that is the canonical stream of consciousness of the subject? If so, does Dennett see these differences? Is one or more of these claims—if they are different—more plausible than the others? Is one or more of these claims better established by Dennett's arguments?
3) Dennett's arguments against 'phenomenological projection' sometimes seem to rest on the premise that there is no (non-metaphorical) Cartesian Theatre: does this viciously beg the question?
4) What exactly is the "metaphysically dubious" notion of "the objectively subjective"? Is this notion really required to make sense of the Cartesian Theatre model of consciousness? Is it really as dubious a notion as Dennett suggests? Is it so dubious as to entail verificationism with respect to the subjective (even if verificationism is not reasonable *tout court*)?
5) On the CT model, for A *to be represented as* experienced earlier than B is for it to *be* experienced earlier than B. Dennett argues against this manner of representing the timing of experiences, but is his rejection of this way of experiencing time inconsistent with our *phenomenology* of time? (What would Dennett say about this?) If Dennett is right about subjective time, can the Cartesian Theatre be rescued?
6) How central is the role of 'autostimulation' in Dennett's account of the evolution of consciousness? How plausible is this account?
7) How literally should we take the notion of a 'meme'? How literally should we take Dennett's claim that human consciousness is, in large part, a collection of memes (meme-effects in brains)? Is it clear how a collection of memes can *constitute* a serial von Neumannesque virtual machine? Does this emphasis on memes mean that non-humans—non-cultural creatures—cannot, properly speaking, be conscious?
8) How, exactly, does the invocation of a serial 'virtual machine' help to explain the emergence of self-control from competing multiple drafts/specialist modules (and, incidentally, are these the same competition or two different ones)? Does the Joycean machine *determine* which 'coalitions' of drafts/modules are victorious at a particular moment, or is it *constituted* by sequences of such victories? If the latter, does it actually have any explanatory role in accounting for self-control? If the former, has Dennett given any indication of how it performs this role?
9) How seriously does Dennett actually take the 'introspective evidence' for Joycean consciousness? Does he adequately explain it?
10) What exactly is Dennett's argument against the Conceptualizer module? How good is it? What follows from its conclusion?
11) Dennett argues that his account of intentional action (including the utterance of speech acts), though it construes them as the mechanical product of 'stupid' sub-processes, does not render these actions non-rational. Is he right about this? What would be the consequences if he were wrong?
12) Is it clear that the Joycean machine, as Dennett has described it, has the powers he attributes to it at the end of Chapter 9? Is it clear—or *prima facie* plausible—that the possession of these powers is necessary and sufficient for consciousness? If not, is this because—as Dennett might argue—we are in the grip of a distorting (and fruitless) ideology?

PART III: THE PHILOSOPHICAL PROBLEMS OF CONSCIOUSNESS

10. Show and Tell

Dennett notes that there is empirical evidence that we manipulate three-dimensional mental images by *mentally rotating them*—that is, it seems that at least some of our mental imagery is at least quasi-pictorial. Does this show that images are displayed and manipulated on an internal screen: is the Cartesian Theatre back? Dennett argues not, using his CADBLIND model to replace the CT one.

First, he notes that the empirical data is not exactly as we would expect if we were literally rotating internal 3D images—we cannot perform tasks (spotting the red X in a complex shape rotation; noticing a sideways Texas; mentally visualising simple crossword puzzles) that, *if* they were imagistic, ought to be no more difficult than the transformations we can mentally achieve.

Then Dennett describes a sequence of 'CAD systems for blind engineers.' Mark I is a CAD system which is 'read' by a *Vorsetzer*—a computer vision system that scans the screen of the CAD system and determines what appears there. In Mark II, the screen and camera are replaced by a cable which carries the information about the bitmap (that would have been displayed on the screen and read by the camera). In Mark III the bitmap itself is replaced by the coding information that the CAD system would have used to create the screen image.

CADBLIND Moral 1: Dennett suggests that Mark III CADBLIND systems will work perfectly only if the CAD-module and the Vorsetzer-module 'speak' the same code—in cases where they do not, the system will have to revert to the Mark II method of sharing bitmap (i.e. image-like) information. He suggests that this is (rather like) what happens sometimes in our brains, especially when we use parts of the brain which detect patterns in visual images. But the experimental examples of imaging failure show that we cannot be solely Mark II systems—we must be (metaphorically) largely Mark III. "People are not CADBLIND systems … but it does prove that we don't have to postulate a Cartesian Theater to explain the human talent for solving problems 'in the mind's eye'" (297).

CADBLIND Moral 2: Whichever version of CADBLIND best resembles our processing of mental images "it is all 'tell' and no 'show'" (296)—for example, in the case of the red X, only in the Mark I system is there any actual redness in the system: in both versions II and III there is only information *about* redness. "All the red is gone—there are only numbers in there" (297).

If thought is not (always) like pictures, is it like language? Dennett claims that fundamentally it is like *neither*: "the media used by the brain are only weakly analogous to the representational media of public life" (303). Some of the 'high-level,' 'temporary' structures of our brain are certainly image-like or linguistic, Dennett contends, but, just as there aren't *really* pictures in the brain, there is no "*language of thought*, a single medium in which all cognition proceeds" (302).

What about our "everyday concept of consciousness" (304)—does this support a kind of CT picture? Dennett uses Rosenthal's analysis of (a central element of) this everyday concept: "[o]ur everyday folk psychology treats reporting one's own mental state on the model of reporting events in the external world" (306).

Two distinctions:

a) To *express* a mental state is to do something that makes manifest that state to an observer. To *report* a mental state is to utter a speech act that *expresses* a belief about that state. (Introspective reports are thus always expressions of *second-order* mental states.)
b) A *belief* is a stable, underlying dispositional state. A *thought* is an occurrent, transient state.

"Since a hallmark of states of human consciousness is that they can be reported … it follows, on Rosenthal's analysis, that 'conscious states must be accompanied by suitable higher-order thoughts, and nonconscious mental states cannot be thus accompanied'" (307, including a quotation from Rosenthal). (Note, of course, that not all thoughts need be consciousness.)

So what happens—on the folk view—when one reports one's mental state (say, a desire that *p*)? I must have a belief about that desire, and this belief must give rise to a thought (the 'occurrent version' of the belief), and this thought is expressed by a speech act.

If this second-order thought needed to be conscious as well, then a vicious infinite regress would loom: there would have to be a (third-order) belief/thought about this thought, and so on. But, according to Rosenthal, the folk do not require that the thought expressed be conscious in order for the thought reported to be so. [This type of theory has since come to be known as the Higher Order Thought—HOT—theory of consciousness.]

Dennett ultimately disagrees with all this as a theory of consciousness, but he endorses it as an account *of our everyday concept of consciousness*. He tries to use it to show that it:

i. "discredits the idea of zombies—with no help from the outside" (304), and
ii. contains internal difficulties that show it should be discarded in favour of an MD-friendly model.

How does it discredit zombies? Suppose we take a crude zombie—which has psychological states but lacks consciousness—and add on self-representation systems: Dennett calls such a system a *zimbo*. "A zimbo is a zombie that, as a result of self-monitoring, has internal (but unconscious) higher-order informational states that are about its other, lower-order, informational states" (310). Such a system—and only such an advanced zombie—could clearly pass the Turing test, and furthermore it (apparently) satisfies Rosenthal's criteria for consciousness. "*It* would think it was conscious, even if it wasn't!" (311). But the notion of a zombie collapses into that of a zimbo, since zombies must be indistinguishable from real persons and so must have the same self-representation capacities as us, so the notion of an 'unconscious zombie' (i.e., a zombie *simpliciter*) appears to be *incompatible* with our ordinary conception of consciousness. Hence, zombies are (according to Dennett) 'discredited.'

What are the internal contradictions of the folk view? It seems to suggest (on the model of ordinary perception) that we introspect by first 'observing' a mental state [1], thereby producing a belief about that state [2], which in turn generates a thought (of that belief) [3], which produces a communicative intention [4], which is then expressed in a report [5]. Between each item in this causal chain there is room for error to creep in (if only when we remember mental events which are just in the past), Dennett argues … and by a move which is unclear to me, Dennett also seems to suggest that this means there must be *contentful states* between each of 1–5 (and then contentful states between each of those, and so on…). [It's possible his argument is that error entails a change of content between the causal stages, and since mental states are 'individuated by their content' this means there must be an additional state in the gap … but since this argument seems clearly unsound I'm reluctant to attribute it to Dennett.] Dennett also suggests that "we end up having to postulate differences that are systematically undiscoverable by any means, from the inside or the outside" (319), which he takes to be a sort of *reductio*.

So, Dennett denies this folk model: instead, on the MD view, our framing of the report *is* the creation of the thought. "We don't *first* apprehend our experience of the Cartesian Theater and *then*, on the basis of that acquired knowledge, have the ability to frame reports to express; our *being able to say* what it is like *is the basis for* our 'higher-order beliefs'" (315). For Dennett, the higher-order state causally depends on the (public or private) expression of a speech act, and not the other way around. On this view, we determine what we are thinking by 'talking to ourselves' *not* by observing ourselves. Furthermore, instead of an ontology of "discrete contentful *states*" we have "a *process* that serves, over time, to ensure a good fit between an entity's internal information-bearing events and the entity's capacity to *express* (some of) the information in those events in

speech" (319). Beliefs, thoughts, experiences etc. *are just heterophenomenological artefacts*, according to Dennett—they do not 'really exist.'

11. Dismantling the Witness Protection Program

In this chapter Dennett continues to hammer away at the idea that there is a 'central witness' (plus the raw phenomenal data it 'witnesses').

Consider blindsight: a neurological condition where patients report an area of the visual field in which they have no conscious visual experience, but in which they can be cued to respond to visual stimuli with a success rate much greater than chance. Does this show, as has sometimes been asserted, that the *consciousness* of vision can come apart from the mere processing of visual data in the brain? I.e., is blindsight partial zombiehood? Dennett says no. The data supporting blindsight are crucially dependent on heterophenomenological data—the patients' reports of their own blindness: it's (nothing more than?) a difference in the way blindsight patients *behave*, compared to, say, those with hysterical blindness. In the rest of this chapter, Dennett seems to suggest that we can add various factors back into blindsight patients until they have full visual consciousness, and that this process of addition need involve nothing suspiciously 'phenomenal.' (Note: this is only a thought experiment: actual blindsight patients may not be trainable in this way.)

a) First, train blindsight patients to increase the accuracy of their responses, using feedback.
b) Then, train them to accurately makes guesses about visual stimuli *without* being cued. This, Dennett suggests, is like being able to *see* something but still not being able to *notice* it (as when a thimble is hidden in full view), so we do not yet have full-blooded intentionality. What is missing is the ability to *notice* the stimulus.
c) Dennett analyses *noticing* on the MD model: as the capacity for contents in the multiple drafts 'pandemonium' to perpetuate themselves long enough to be reported. This can be achieved by training he suggests, as in the example of the piano tuner. And *this*, Dennett suggests, is full-blown visual consciousness.

But haven't we failed to add back in the *phenomenal properties*—the qualia—of visual consciousness, Otto asks? There's no such thing, Dennett replies—there *are no* "special properties or features of our experience" with which we are "*directly acquainted*" (359). There is no "*actual* phenomenology" (365).

a) If a 'blindsight' patient became behaviourally more or less just like a normally sighted person, we would no longer take seriously any heterophenomenological reports they might make of their own lack of visual consciousness.
b) The notion of sensory qualia is closely related to the notion of 'filling in,' and Dennett argues that 'filling in' is a myth: there are no mental 'colours' or 'sounds'—what Dennett calls *figment*—that are used to 'fill the gaps' in experience.

Instead of there being actual colours in the brain, Dennett points out, the brain (vector) *codes* for colours. Furthermore, according to Dennett, the brain *never decodes* this representation, and puts the colours back in—the codes are all there are, neurally speaking. [To think otherwise is what he has called, elsewhere, "the myth of double transduction."] Dennett's explanation of 'filling in' is simply that the brain codes for *regions* (somewhat like the numbers on a colour-by-numbers picture), rather than by compiling a complete bitmap of the visual field. In the absence of any contradictory visual information, the brain 'automatically' codes certain areas of the visual field in certain ways: e.g. a clear look at some small areas of wallpaper is sufficient to result in a coding for the whole expanse.

Consciousness, Dennett claims, is massively gappy and discontinuous: it's not a 'plenum.' The reason we do not notice this is that we have no 'epistemic hunger' for the missing information—e.g. information missing in

the blind spot or between saccades—and so do not notice it. "The fundamental flaw in the idea of 'filling in' is that it suggests that the brain is providing something when in fact the brain is ignoring something" (356). "[T]he absence of representation is not the same as the representation of absence. And the representation of presence is not the same as the presence of representation" (359). (Different kinds of neurological neglect can thus be explained, Dennett suggests, as "a pathological loss of epistemic appetite" (356).)

So, if we do not 'fill in' a representation of the colour patch missing in our blind spot, Otto asks, then *where* is the colour that we see? Dennett's answer is that it is out in the world—for any point on the visual field, whenever we want to check that it's a high-resolution plaid then we look there and see that it is so. This is what Minsky has called the Immanence Illusion: "Whenever you can answer a question without a noticeable delay, it seems as though that answer were already active in your mind" (Minsky, quoted on 360).

Dennett usefully summarises the plot so far on pp. 362–368.

12. Qualia Disqualified

Dennett argues that the notion of qualia is so conceptually tangled that it is best if we just jettison it entirely. Instead, 'the way it is with me' is nothing more than "the sum total of all the idiosyncratic reactive dispositions inherent in my nervous system as a result of my being confronted by a certain pattern of stimulation" (387).

Dennett diagnoses the core intuition behind qualia in terms of secondary qualities: they are putatively (intrinsic properties of) the things produced in the mind by dispositional properties like colour, sound, smell etc. That is, we think, the 'idea of red' is not just *about* red—it *is*, in some sense, (occurrent) red. "[I]f there is no inner *figment* that could be coloured in some special, subjective, in-the-mind, phenomenal sense, colors seem to disappear altogether! *Something* has to be the colors we know and love, the colors we mix and match. Where oh where can they be?" (370–371).

This intuition is just what Dennett denies, and hopes to replace. There is no figment: there are only *colour judgements*. On Dennett's view, colours are light-reflecting properties of objects out in the world; these properties "cause creatures to go into various discriminative states, scattered about in their brains, and underlying a host of innate dispositions and learned habits of varying complexity" (372). That is, these discriminative states *themselves* have secondary/dispositional properties (which Dennett calls 'playing Locke's card a second time'), such as disposing humans to make verbal reports of colour discriminations. These dispositional properties of discriminative states *are sufficient to explain the heterophenomenology of colour* (according to Dennett)—no 'extra' (intrinsic, subjective, private, ineffable) qualia are necessary or philosophically desirable. E.g. Dennett's account—he says—can explain how we make comparative colour judgements (in a CADBLIND-ish sort of way).

Colours are an interesting case because it is very complex to say what 'objective' properties in the world colours might be [and this tends to lead people away from colour realism … which tends to lead people to think that the subjective experiences of colour must defined *prior to and independently of* their causes … Dennett argues against this]. "[T]here is no simple, nondisjunctive property of surfaces such that all and only surfaces with that property are red" (376). But, Dennett argues, instead of puzzling over how we could detect such weird, pre-existing properties in our environment, it is evolutionarily plausible to hold that colours and our colour-detecting mechanisms were 'made for each other.' "It is a mistake to think that first there *were* colours … and then Mother Nature came along and took advantage of *those* properties by using them to color-code things. It is rather that first there were various reflective properties of surfaces, reactive properties of photopigments, and so forth, and Mother Nature developed out of these raw materials efficient, mutually adjusted 'color'-coding/ 'color'-vision systems, and among the properties that settled out of that design process are the properties we normal human beings call colors" (378). There are no 'colours' specifiable

independently of an observer-class. (And thus, before observers, there were no colours specifiable at all … except the infinite number of 'colours' relative to *merely possible* observers. This is what Dennett means when he says that secondary qualities are *lovely*—defined with respect to an observer-class (rather than *suspect*—which entails that something has already had a particular effect on a particular observer). That is, they are in some sense subjective or relative, rather than objective—though no less explicable in evolutionary terms.)

This story, Dennett suggests, shows why qualia are 'ineffable': we cannot define the property *M* any better than by saying it is the property detected by *this* M-detector. "Otto points to his discrimination-device, perhaps, but not to any quale that is exuded by it, or worn by it, or rendered by it, when it does its work. There are no such things" (383).

"But still [Otto insists], you haven't yet said why pink should look like *this*! … Like the particularly ineffable, wonderful, intrinsic pinkness that I am right now enjoying" (383). Dennett chooses to focus on the word *enjoying* in this imagined objection, and—to 'accommodate' this intuition—he describes how sensations originated in signals that have an *affective* component—as either 'warners' or 'beckoners'—and that our modern perceptions are still constructed (in highly complex ways) from these basic mechanisms.

"What qualia *are*, Otto, are just … complexes of dispositions. … That 'quale' of yours is a character in good standing in the fictional world of your heterophenomenology, but what it turns out to be in the *real* world in your brain is just a complex of dispositions" (389). The difficulty that we have in knowing 'what it is like' for other people is simply the practical difficulty of making our reactive dispositions the same as theirs (e.g. the same as those of a Leipzig Lutheran churchgoer in 1725).

One common response to this kind of assertion about qualia is to appeal to the possibility of inverted qualia: it seems to be (logically) possible to invert one's qualia while leaving all one's reactive and associative dispositions totally unchanged (e.g. if we exactly inverted the subjective colour spectrum). If this is so, then qualia must be something *over and above* mere dispositions.

Dennett objects that—in principle—there would never be any intersubjective way to verify that such an inversion had taken place: this difference in qualia must be a difference that makes no difference at all, since it *ex hypothesi* has no effect on functional or behavioural dispositions. Although commonly recognised by qualophiles, this fact does, according to Dennett, "provide support … for the shockingly 'verificationist' or 'positivist' view that the very idea of inverted qualia is nonsense—and hence that the very idea of qualia is nonsense" (390).

But wouldn't *intra*subjective qualia inversions be verifiable? Merely 'switching the qualia' will not work, since this will alter the subject's reactive dispositions. So we must add an additional 're-wiring' "*after* the inverted qualia have taken their bow in consciousness, but *before* any of the inverted reactions to them can set in. But is this possible? Not if the arguments for the Multiple Drafts model are correct. There is no line that can be drawn across the causal 'chain' from eyeball through consciousness to subsequent behavior such that all reactions to *x* happen after it and consciousness of *x* does not happen before it" (392).

Dennett concedes (for the sake of argument) that we could imagine all of the subject's reactive dispositions gradually *adapting* back to normal after a surgical qualia inversion, but argues that it is then *indeterminate* whether your qualia remain inverted or not (bearing in mind that we are no longer permitted to classify all adaptations as either Stalinesque or Orwellian). In fact, Dennett seems to suggest, the most natural explanation is that the renormalisation of the dispositions *constitutes* a re-inversion of the 'qualia'. E.g. consider the case in which people learn to like the taste of beer: does the way the beer tastes to them change, or do they come to like the taste it always has? According to Dennett, there is no principled way to answer questions of this type: all we can do is "decide to *reduce* 'the way it tastes' to one complex of reactive dispositions or another. … So if a beer drinker furrows his brow and gets a deadly serious expression on his face and says that what he is referring to is '*the way the beer tastes to me right now*,' he is definitely kidding himself

if he thinks he can *thereby* refer to a quale of his acquaintance, a subjective state that is independent of his changing reactive attitudes" (396).

Hence qualophiles, according to Dennett, are just unable to produce an obvious case in which qualia come apart from reactive dispositions. The idea that this is possible is, he says, simply fall-out from the attractive myth of the Cartesian Theatre.

A second common kind of argument from the qualophiles is what has become known as the 'knowledge argument,' especially Jackson's thought-experiment about Mary, a brilliant colour scientist who has nevertheless never seen colour before in her life. Despite knowing everything there is to know about the physics of colour, she will still—Jackson asserts—learn something surprising and new the first time she sees colour. There must, therefore, be colour facts (about, presumably, qualia) over and above all the physical facts.

Dennett objects that this thought-experiment only seems convincing if we "are simply not following directions!" (399). If we *really* imagine that Mary knows *everything* physical about colour then, for example, she will not be tricked by a blue banana (since she knows the usual colour of bananas, and knows what effect such a colour surface should have on her nervous system, and so on). As such, it is simply *not clear* that she will learn anything new. "Jackson's example … is a classic provoker of Philosophers' Syndrome: mistaking a failure of imagination for an insight into necessity" (401).

One consequence of Jackson's thought experiment, if it were successful, would be that qualia are epiphenomenal (since, roughly, they are not physical and only the physical has causal powers). Dennett points out that the term 'epiphenomenal' is ambiguous. The root meaning of the term is for something that is *non-functional* with respect to some process or other (e.g. the hum of a computer)—they are by-products, but they are still products with lots of effects on the world. The *philosophical* meaning of the term, by contrast, is that something "is an effect but itself has no effects in the physical world whatever" (402). But the philosophical meaning, Dennett argues, is too strong: there could never be any empirical reason at all to assert the presence of an epiphenomena, since by definition they are undetectable by any possible (physical) measuring instruments and have no causal effect whatsoever on behaviour (including the behaviour of people who report their own epiphenomenal qualia). To appeal to *first person* evidence for qualia leads one—according to Dennett—to solipsism, since this evidence itself (beliefs etc.) must be totally cut off from the physical world. And, says Dennett, there are no good *a priori* reasons to believe in epiphenomenal qualia. So to say "that genuine consciousness is epiphenomenal *in the ridiculous* [philosophical] *sense* … is just silly. … I can't prove that no such sort of consciousness exists. I also cannot prove that gremlins don't exist. The best I can do is show that there is no respectable motivation for believing in it" (405–406).

So—according to Dennett—it is just "a woebegone mistake" to assert the existence of epiphenomenal qualia in the philosophical sense. And 'qualia' which are epiphenomenal in the other sense are no threat to materialism (or the MD model).

13. The Reality of Selves

Are there really selves? Obviously *we* exist. But, just as obviously, there are no self-entities *over and above* our brains. Things with selves evolved from things without selves, so there has to be a true story about how selves came to be: Dennett suggests the earliest stages of this story comes with the evolutionary emergence of *boundaries* between 'inside' and 'outside.' "[E]ven such a simple self is not a concrete thing but just an abstraction, a principle of organization" (414). Dennett gives a sequence of examples designed to show how biologically fundamental this principle is, and how the creation and extension of these boundaries (e.g. a termite's nest) can take place according to 'mechanical' biological principles.

Dennett then extends this idea to cultural, human methods for extending boundaries—e.g. clothes, cars—and

suggests that the most important difference between human boundary-making and that of other species is that we use *language*. We not only use words to represent ourselves to the world, but we also use them to represent ourselves to ourselves—to self-represent. "Our fundamental tactic of self-protection, self-control, and self-definition is not spinning webs or building dams, but telling stories, and even more particularly concocting and controlling the story we tell others—and ourselves—about who we are. And just as spiders don't have to think, consciously and deliberately, about how to spin their webs … we … do not consciously and deliberately figure out what narratives to tell and how to tell them. Our tales are spun, but for the most part we don't spin them; they spin us. Our human consciousness, and our narrative selfhood, is their product, not their source" (418).

The *self*, then, is the "center of narrative gravity" (418) of these stories: it is a theoretical fiction, postulated in order to immensely simplify heterophenomenology. A key idea, for Dennett, is that the stories are *not* spun by a "central chief executive" (420) (which then would be a candidate to be 'the real self')—instead, like a termite mound, selfy narratives are (presumably) produced by much simpler, mechanical processes. Selves are not "brain-pearls" (424).

Dennett denies that, for selves, "it must be All or Nothing and One to a Customer," and tries to back this up with cases from Multiple Personality Disorder, identical twins who apparently share a single self, 'gappy' selves, and split-brain patients. The continuing existence of selves depends on nothing more or less than the persistence of a narrative.

The Centre of Narrative Gravity "plays a singularly important role in the ongoing cognitive economy of [its] living body, because, of all things in the environment an active body must make mental models of, none is more crucial than the model the agent has of itself" (427). E.g. (according to Dennett) we have to have 'wired in' knowledge of which thing in the world we are so that, for example, we do not eat ourselves. Sometimes, the inputs we receive from our environment are more puzzling, so that we need to be able to *work out* which thing is us, e.g. by doing something and seeing what moves. We also need to be able to represent to ourselves "our own internal states, tendencies, decisions, strengths, and weaknesses" (428), and we do this in a similar fashion: by telling stories to ourselves about these things, and then checking them for accuracy.

"And where is the thing your self-representation is *about*? It is wherever you are. And *what* is this thing? It's nothing more than, and nothing less than, your center of narrative gravity" (429).

14. Consciousness Imagined

Dennett points out that it is a consequence of his theory that there could be silicon-based conscious robots. It's hard to imagine *how* a 'bunch of silicon chips' could be conscious … but "it's just as difficult to imagine how an organic human brain could support consciousness. … It turns out that the way to imagine this is to think of the brain as a computer of sorts. … By thinking of our brains as information-processing systems we can gradually dispel the fog and pick our way across the great divide, discovering how it might be that our brain produces all the phenomena. … The huge gap between phenomenology and physiology shrinks a bit; we see that some of the 'obvious' features of phenomenology are not real at all: There is no filling in with figment; there are no intrinsic qualia; there is no central fount of meaning and action; there is no magic place where the understanding happens" (433–434).

There is thus, Dennett says, no reason to succumb to philosophers who argue that there is *in principle* no way to understand how the organic brain can produce consciousness. Arguments in favour of this position either ignore arguments like Dennett's, or depend on thought experiences that "dissuade the reader from trying to imagine, in detail, how software could accomplish [consciousness]" (435), such as Block's Chinese Nation and Searle's Chinese Room. Such arguments tend to work, Dennett suggests, by noting that there is no 'genuine understanding' in the simple components of the system, and "[t]hen comes the suppressed premise: Surely

more of the same, no matter how much more, could never add up to genuine understanding. But why should anyone think this is true? … [I]f … we are materialists who are convinced that one way or another our brains are responsible on their own, without miraculous assistance, for our understanding, we must admit that genuine understanding is somehow achieved by a process composed of interactions between a host of subsystems none of which understand a thing by themselves" (438–439).

How about Nagel's claim that we can, in principle, never know what it is like to be a bat? Dennett claims we need to distinguish between the *epistemological* question of knowing what would count as having experiences like a bat and the more *practical* question of whether we could ever transform (part of) our minds (temporarily) into bat minds. It is the former question which is the key one, and Dennett 'flatly denies' that no amount of third-person data could ever put us in that epistemological position: he argues that if we collect enough information about bat sensory systems then, for example, we will be able to tell the difference between plausible and implausible heterophenomenological narratives about them. Dennett notes, however, that while cognitive ethology allows us to 'read off' a kind of heterophenomenology from animal behaviour, non–language-using animals like bats have simple selves but not "selfy selves … no Center of Narrative Gravity, or at most a negligible one" (448).

Do conclusions like this make Dennett's theory somehow immoral? He (of course) argues not. In fact, theories like his allow us to *explain* 'mattering'—suffering and enjoyment—in terms of the satisfaction or frustration of cognitive processes. ("The idea of suffering being somehow explicable as the presence of some intrinsic property—horribility, let's say—is as hopeless as the idea of amusement being somehow explicable as the presence of intrinsic hilarity" (449). To cling to the importance of *intrinsic consciousness per se* as the source of mattering is to "cling to doctrines about consciousness that systematically prevent us from getting any purchase on *why* it matters" (450).

A few questions for critical thought

1) When we imagine red things (say, the Canadian flag) is it really true that "[a]ll the red is gone—there are only numbers in there"? What might Dennett say about this example?
2) Why does Dennett deny that there is a "*language of thought*, a single medium in which all cognition proceeds"? How good are his reasons for asserting this?
3) Is it really "a hallmark of states of human consciousness … that they can be reported"? How significant is this assumption for Dennett's theory of consciousness? Is the HOT account really a good story about the 'folk' understanding of consciousness?
4) Are we all zimboes?
5) How adequate is Dennett's rejection of the HOT view of consciousness?
6) Are there (in *any* sense) colours in the brain? If not, must Dennett be right about colour qualia? When Otto asks "where are the colours that we see?" is Dennett's answer adequate?
7) How heavily does Dennett lean on the suppressed premise that any account which *is sufficient to explain the heterophenomenological data* is sufficient to explain *consciousness*?
8) Does Dennett beg the question against inverted qualia by appealing to the MD model in his 'refutation' of the possibility?
9) How effective is Dennett's example of the beer drinker against the notion of reliably internally ostending our own qualia?
10) Does Dennett conclusively rule out the theoretical possibility of epiphenomenal qualia? If so, what impact does this have on the debate about consciousness?
11) What exactly *is* a 'centre of narrative gravity'? Does Dennett really mean (as he at one point says) that the narratives we spin are somehow *about* our centre of narrative gravity?
12) In what sense, according to Dennett, can we know what it is like to be a bat? Will this satisfy Nagel?