Stage Effects in the Cartesian Theater: A review of Daniel Dennett's Consciousness Explained

Kevin B. Korb
School of Computer Science and Software Engineering
Monash University
Clayton, Victoria 3168
Australia

korb@bruce.cs.monash.edu.au

PSYCHE, 1(4), December 1993
http://psyche.cs.monash.edu.au/v1/psyche-1-4-korb.html

Keywords: philosophy of mind, qualia, functionalism, multiple drafts model, Cartesianism, folk psychology

1. Dennett's Multiple Drafts Model of Consciousness

1.1 `We're all zombies. Nobody is conscious' (Dennett 1991, p. 406) is an assertion Dennett is actually brought to make in his attempt to maximally provoke his readers; his title serves a similarly provocative role.<1>

1.2 Judging by such expressions, Dennett undoubtedly would like his work to be even more provocative than it actually turns out to be. The main thrust of *Consciousness Explained* is to apply a widely accepted thesis about the relation between mind and brain---non-homuncular functionalism---in a program of philosophical therapy to rub away a variety of puzzles raised by both philosophers and experimentalists about consciousness. I believe that the central thesis will be relatively uncontentious for most cognitive scientists, but that its use as a cleaning solvent for messy puzzles will be viewed less happily in most quarters. I will start by briefly presenting Dennett's views on functionalism and method and then present some concerns about Dennett's sanitation projects.

1.3 Functionalism is the theory that evolved out of mind-brain identity theory, the thesis that mental states or processes are just identical to some particular brain states or processes. That theory, introduced by J. J. C. Smart and U. T. Place in the 1950s, suffers from the objection that mental states cannot be shared since the physical goo of our

brains cannot be shared. Most obviously, Martians cannot have pain states if they do not have brain states (i.e., their `brains' may be made out of something other than neural goo; cf. D. Lewis, 1980); but even more radically, you and I cannot have the same pain states since you and I cannot share brain states. Theorists, in response to the pressure of such arguments, have tended toward functionalism, or the token-identity theory: mental states are to brain states as types are to tokens---as say having one dollar is to having a particular dollar bill (or to having a Susan B. Anthony coin). Although mental states are not identical to brain states then, they *supervene* upon brain states; that is, pain (or whatever) is realizable via some kinds of brain state. This does not rule out other possible varieties of realization; for example, Martians may realize a pain state by way of some fluid hydraulic system, rather than a neural system. What identifies it as a *pain* state is not the physical token state that constitutes it, but the functional (more generally, causal) relations which that token state has to other token hydraulic states and to the system's overall behavior. Dennett, together with most cognitive scientists (but certainly not all), is a believer in some flavor of functionalism.

1.4 He is also opposed to homuncular theories, again together with most commentators. A homuncular theory is one that proceeds to explain some cognitive (conscious) ability or process by first producing some functional analysis that appears to make progress for a time, but then, when things get messy, simply invents an internal agent---a homunculus---which turns out to have all the abilities that needed explaining in the first place. Since the homunculus in its turn will need to be explained, there is an immediate threat of infinite regress. The most extreme variety of homuncularism is the dualist ghost in the machine: no matter how good an explanation of mental *function* you may produce in terms of brain processes, a committed dualist will complain that it is incomplete because it fails to explain that aspect of mentality which is non-physical. If someone offers a functional account of natural language understanding, for example, a homuncular dualist will insist that in addition to all of *that* (whatever it may be), there must be a creature in the middle of it all that has the *real* understanding, namely the Mind. Neither dualism nor homuncular functionalism are sufficiently appealing to have many adherents nowadays: where writers are moved to issue an explicit opinion, they are very largely negative opinions.

1.5 Despite this, Dennett shows that the homuncular concept retains a powerful grip on the imaginations of many, perhaps most, cognitive scientists. While explicit dualism and homuncularism are (no doubt properly) `endangered' theses, a great many theories and judgments advanced by cognitive scientists rely at some point upon there being a magical place in the head where everything comes together---in what Dennett calls the Cartesian Theater. This concept is pernicious in a variety of ways. For one thing, it leads to lazy analysis: if we can rely upon some arbitrarily complex central process to clean up our functional loose ends, we needn't be very careful about specifying whatever functional processes we do provide. But worse, this Cartesian Materialism (functionalism with the Theater at the center) again leads to infinite regress: if there is a theater where consciousness is `projected', then there must be an observer viewing the projection (else why bother with the theater?). As before, we will find it difficult to understand this observer: if the theater and its audience are needed to understand conscious processes,

then an `inner' theater and `inner' audience will be needed to understand the observer, and so on. But if the theater and its observer are not needed to understand conscious processes, then why introduce them in the first place? As Dennett notes, the best place to stop an infinite regress is usually at the beginning.

1.6 It is not so difficult to imagine that the brain---and consciousness---might be analyzable into a set of processes with no dominant center. Certainly it is not so difficult to build computer programs which behave in complex ways, including learning and adapting to a complex environment, without there being a `central authority' which decides on every action: Oliver Selfridge provided a model for this in his program Pandemonium in the 1950s (Selfridge, 1959). In similar spirit, Dennett proposes in opposition to Cartesian Materialism his Multiple-Drafts model of consciousness: the idea that consciousness is more or less a continuing (but not continuous) flow of drafts of text; the texts are constantly being edited and re-edited, passed from one process to another. Sometimes they initiate speech, sometimes they are stored in long-term memory, other times they are entirely forgotten.

1.7 What counts as the content of consciousness on this model is by and large up for grabs. But even such vagueness may be appropriate when we reflect that we do not have any clear idea of what to count as the content of consciousness independent of our theories of consciousness. There is one criterion for such content that has been widely accepted among investigators: what a subject *reports* expresses some conscious content. This is, as Dennett remarks, what we *normally* (but not always) regard as sufficient evidence for a human to be, or have been, conscious of the content expressed. This gives us a counterfactual criterion of consciousness, which we may call the `probe criterion': if an agent were to inquire of the subject what she or he was aware of (or whether she or he was aware of X), then if the subject would have reported being aware of X then the subject was conscious of X. This criterion, however incomplete, is useful. Indeed, Dennett uses it to point out that an example very often employed by philosophers to argue that certain complex cognitive tasks are performed unconsciously (e.g., by Dennett himself, 1969; and by me in Korb, 1991) is unsuccessful: in `automatic driving' people appear to *have been* unaware of their `actions' for a few minutes while driving and simultaneously day-dreaming. But Dennett notes that if someone *were to have* inquired about say the bridge just crossed *during* the reveree, the driver would very likely *not* respond `What bridge?' Automatic driving is better explained by rapid forgetting than by unconscious behavior.

1.8 The adoption of this criterion does not mean that Dennett is (or should be) passively acceptive of the *content* of subjective reports *about* the contents of consciousness. Introspective reports are not random mumblings: they *express* some conscious content; but at the same time they may *report* a quite different content. The most obvious example is a lie about one's own beliefs: such a lie reports one content (what you would have your interlocutor believe) while expressing another (your internal belief state). The larger problem with introspective data has been the ease with which people confabulate mental events that apparently `aren't there' at all. But this and related difficulties do not imply that introspective reports have no scientific value, as some extreme behaviorists would

have it. The behavior of making such reports is itself as factual as any behavior scientists are otherwise called upon to explain. The content of introspective reports typically describes---in what Dennett calls the subject's heterophenomenology---a `mental world' that is more or less coherent, bears certain relations to what is known about neuroscience, and has some pattern of similarities and dissimilarities with the introspective reports of others. What cognitive science should be aiming at is producing an explanation of these facts that is concise, that covers as much of the data as possible, and that stands up well to futher tests. This is just, of course, what we demand of any scientific theory. And we should expect that the resultant theory will end up *denying* the face-value interpretation of some kinds of introspective reports, even when those reports are repeated by all comers. This is again strictly in parallel with other sciences: almost all of our theories describe some cases where appearances are simply deceptive (e.g., almost every measurement of a star's location will be misleading within some small angular distance from the sun).

1.9 I have summarized the main elements of Dennett's Multiple Drafts model of consciousness and the methodology he proposes that the cognitive sciences adopt for its study. While not uncontroversial nor unoriginal, I expect thus far the story is congenial to most functionalists. Less so will be Dennett's applications of the theory to various puzzles about consciousness: despite suggestions to the contrary, they do not simply fall out as corollaries to the Multiple Drafts model. A few examples follow.

2. The Phi Phenomenon

2.1 The `phi phenomenon' refers to the apparent motion created by rapidly presenting a series of still pictures each of which shows a slight displacement of some image---the basis of motion pictures. This is an interesting psychological effect made even more interesting in an experiment proposed by Nelson Goodman and conducted by Kolers and von Gruenau (1976): when lights are successively illumined in a series (as in some neon signs) we seem to see a single light in motion; what would happen if two successively illuminated spots were of different colors, say green and red? What would happen to the color of `the' spot as it moved from the first to the second position? The experimental answer (p. 114): `The first spot seemed to begin moving and then change color abruptly *in the middle of its illusory passage* toward the second location,' which it reached 150 msec later (when the red spot was illuminated). This raises a few strange questions: how can the apparent colored spot move in the right direction, toward the second location, *before* we have any idea (input) of where the second location is? And, how can the green spot change to red in the *middle* of this traversal before we have any idea (input) that red is what it is supposed to become?

2.2 Dennett discusses two possible models accounting for this phenomenon: Orwellian and Stalinesque revisionism. The Stalinesque model supposes that there is some pre-conscious editing of the input going on. In particular, the green image first received is held up in the editing room until the red image becomes available. The pre-conscious editor then rapidly paints some appropriate intermediate images, splices it all together and passes the result on to the Cartesian Theater for presentation to Consciousness. The

Mind only sees what the Stalinesque editors want It to see. The Orwellian model operates in reverse: Consciousness receives the green image; then It receives neutral background images; subsequently Consciousness receives the red image. The Orwellian editor, realizing the discrepancy, quickly revises the operative history of the episode so that there are appropriate intermediate states. When Consciousness needs to review history, say to answer some question, It finds that the green spot changed to red during an intermediate traversal of the two locations, and reports that as fact. Now, of course, the Orwellian history isn't *written down*; it just *seems* to Consciousness that such a color change was actually *seen*. And what happened to the initial intermediate images that showed no such traversal or color change? They have simply been lost, forgotten in the shuffle.

2.3 So which account sits better with the Multiple Drafts theory? To make a fair assessment we have to eliminate the references to `Consciousness' and `Mind'; but neither account strictly requires them anyway. Dennett claims that there is then nothing to choose between the two models: when we take a sufficiently low-level view of what's going on in the brain, the where and when of consciousness becomes blurry, and so there is nothing to choose between a piece of retroactive Orwellian editing and proactive Stalinesque editing. As there is no `finish line' (Cartesian Theater) against which to measure whether the editing happens before or after Consciousness, there can be no fact of the matter about which of the Orwellian or Stalinesque model is the true one: `there is really only a verbal difference between the two theories' (p. 125). Indeed, there is no fact of the matter about when *anything* becomes conscious.

2.4 Here I think Dennett is simply wrong. Recalling the probe criterion of consciousness, we can ask what the subject would report were a query inserted between the green and red stimuli. Would anyone report that the spot was in-between the two end points? Such a response would require more than retroactive editing, it would require retroactive *causation*! That is out of the question, but it doesn't follow that the green flash, or neutral background, would be reported: perhaps they would have been held up in an editor's room. Of course, we cannot in any case squeeze either an inquiry or a response inside a 150msec window, factually or counterfactually. But the verbal probe was always understood to be a crude device: we can surely suppose that our low-speed probes and their low-speed responses may come to be known to be highly correlated with some specific brain processes which are themselves high-speed processes. Let's simplify and suppose that low-speed reports of green sensations are highly correlated with some high-speed brain process Q. If Q is then observed during these tests before 150msec has elapsed, then this is clear evidence in favor of the Orwellian model and against the Stalinesque model in the same way that the probe criterion provides evidence against the idea that automatic driving is an unconscious process. It is simply false that the two models are merely `verbally different' from one another: the confusion of an inability to test one theory against another *now* with an *in principle* inability to test between them is a fundamental confusion. It is the same as confusing operationalism <2> with the demand that new theories must make *some possible* difference in our observations, which is prerequisite to there being *any possible experimental science* deciding between those theories. (Dennett is similarly confused about verificationism; cf. p. 403.)

2.5 The idea that there is no `when' to consciousness is also wrong. The blurriness of boundaries is no good argument for the non-existence of boundaries. There is no clear boundary between baldness and hairiness, but that is no consolation to those who are bald and don't like it. So long as there is no Consciousness Cell---some specific neuron that fires when and only when a subject is conscious (and who could believe that?)---then there will be *some* vagueness about where and when conscious processes are occuring. But if conscious processes indeed supervene on physical processes, there is nonetheless a where and when to consciousness: all physical processes occur in space-time.<3>

3. Folk Psychology

3.1 Dennett has previously argued that consciousness, whatever it may be, is so different from our ordinary conception of it that we ought to abandon the concept (Dennett, 1979). Here, while not repeating that argument, he claims that the ordinary notions of belief, desire, etc.---the mental predicates of `folk psychology'---ought to be abandoned: `we may use the oversimplified model of folk psychology as a sort of crutch for the imagination when we try to understand self-monitoring systems, but when we use it, we risk lapsing into Cartesian materialism' (p. 320). The eliminative materialists (e.g., Churchland 1981) have similarly argued that `folk' concepts of psychology will someday be replaced by some as-yet-unconceived clean, scientific concepts of neuroscience. Churchland however, had no specific objections to `believe', `want', and so on other than his prejudicial *conviction* (yes: belief) that these terms would not show up in the not-yet-conceived laws of future neuroscience. Dennett, on the other hand, has a specific, but mistaken, objection to `belief'.

3.2 Dennett presents, sympathetically, David Rosenthal's analysis of consciousness in terms of belief (Rosenthal, 1986, 1990). We may *express* our beliefs quite unconsciously, without our knowing that we do so, indeed without our knowing that we have such beliefs (such as some racists who sincerely believe that they are not); but we cannot *report* our beliefs (sincerely) without also being aware of those beliefs. If we call occurent beliefs (those we are aware of *now*) thoughts, in order to report a belief we must be thinking about it; that is, we must have an occurent, second-order belief about the belief being reported. And now (p. 307):

> Since a hallmark of states of human consciousness is that they can be reported (barring aphasia, paralysis, or being bound and gagged, for instance), 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'' (1990, p. 16). The higher-order thought in question must of course be about the state it accompanies; it must be the thought that one is in the lower-order state (or has just been in it---time marches on). This looks as if it is about to generate an infinite regress of higher-order conscious states or thoughts, but Rosenthal argues that folk psychology permits a striking inversion: *The second-order thought does not itself have to be conscious in order for its first-order object to be conscious*. You can *express* a thought without being conscious

> of it, so you can express a *second-order* thought without being conscious of *it*---all you need be conscious of is its object, the first-order thought you *report*.

3.3 The difficulty Dennett has with this analysis is that an explosion of higher-order beliefs may occur nevertheless. On the way to expressing the second-order thought there might creep in some error, so that what is reported is not the first-order belief. That is a relatively mundane kind of error. But why can't there also be an error that intervenes between the experience and the belief about it; can't one's belief about one's current conscious state be mistaken? It seems that we can cut through such complicating possibilities by asserting that reflexive second-order belief implies first-order belief: I believe that I believe P logically implies I believe P.

> But this merger will not quite do the work that needs to be done.... Even if it is intuitively plausible that you cannot be mistaken about how it *is* with you right *now*, it is not at all intuitively plausible that you cannot be mistaken about how it *was* with you back *then*.... The logical possibility of misremembering is opened up no matter how short the time interval between actual experience and subsequent recall---this is what gave Orwellian theories their license. But as we saw in chapter 5, the error that creeps into subsequent belief thanks to Orwelllian memory-tampering is indistinguishable---both from the outside and the inside---from the error that creeps into original experience thanks to Stalinesque illusion-construction. (pp. 318-9)

3.4 Of course, we saw no such thing. Nor does the simplifying principle of my believing that I believe that P implying I believe that P require adoption of the demonstrably false: I believe that I *believed* that P implies that I believe that P. Time is truth. And it is not a *difficulty* of `folk' psychology that we can meaningfully talk of higher- and still higher-order beliefs that do not immediately collapse into first-order belief: that is rather an aspect of the expressiveness of ordinary psychology. Children will soon tire of extending the sentence `I know that you know that I know that you know that ... you want the last cookie', but that's hardly because anyone believes each extension has the same meaning as the sentence it embeds. Rosenthal's analysis has content and merit that is neither exhausted nor explained by Dennett's Multiple Drafts model.

4. Zombies and Robots

4.1 For many philosophers the crux of the matter of consciousness comes down to what to make of qualia---the phenomenal qualities of the things of which we are conscious, the raw feels and sensations that make up much of our conscious lives. One way to get at the issue is to ask, could there be a (philosophical) zombie? That is, could there be something which is physically indistinguishable from ordinary humans but which fails to be conscious? The idea that there *could* be a zombie appears to arise from a few plausible considerations. If the functional role of pain, for example, is to inform us of injury and to motivate us to avoid such injury in the future, there seems to be no reason to disbelieve

that such a role could be occupied by some non-conscious processes. If we had an incredibly complex program for a robot which provided for every human-like cognitive function except pain, then surely we could go on to provide damage-detection sensors and a complex of goals and goal-oriented programming to avoid activating such sensors. The same must be true of all other kinds of qualia. But if we can in principle program such functions, then we can in principle reconstruct that program in neural matter. In that case we have a zombie: for by stipulation we have implemented the function of every conscious phenomenon without the consciousness.

4.2 Dennett does not go along with this; nor can any functionalist. One method he uses to break down this kind of reasoning is based upon the Multiple Drafts model of consciousness. If consciousness is constituted of some ever-changing flux of cognitive processes, some of them operating on the contents of others, then we have only a very fuzzy border between the presence and absence of consciousness. If we think of any one function or small clump of functions in isolation---say, pain functions---then it is easy to imagine implementing those functions in a system without consciousness. However, when we pile one function on top of another on top of another into some extremely complex whole, our intuitions about such matters dissolve. Consider that no single neuron is conscious; nor will it become conscious when connected to *one* other neuron. Nevertheless, if we repeat the process billions of times we must *end up* with consciousness: we know in advance that something like human brains have consciousness. Surely this argument is right: if consciousness is, loosely, a property of the brain as a whole, then it is just ill-conceived to look for it also within the constituent parts of the whole---whether the brain is looked at as a complex of neurons or as a complex of functions.

4.3 Since robots, on the initial argument above, have the same functions as their neural doppelgaengers, they must also have the same functional properties, such as consciousness. So Dennett in fact applies just the same argument to robots, claiming that adding a piece of computational function here to one there may in fact get you to robot consciousness in the end (pp. 438f). The argument here is a bit weaker than that for humans. Since we know in advance that brains are conscious, we can be confident that adding neural functional complexes together can in principle give us something conscious. We know nothing in advance about robot consciousness, however. And certainly without knowing in advance that the goal is achievable the argument goes nowhere: I can always jump just a tiny bit higher than my last jump; this does not mean that after some unimaginably large number of jumps I can reach the moon. Nor is doubt here necessarily just an expression of prejudice against robots: although there are good reasons to believe that computers can capture the full range of *computational* abilities of the human brain, this does not imply that such computer systems will also (necessarily) implement the full range of brain functions---what has not been shown is that all important brain functions are strictly reducible to computations.

4.4 Qualia, on Dennett's view, are neither more nor less than dispositional properties of cognition. The idea that there is something mysteriously ineffable about consciousness is particularly his object of attack. If there is some content of consciousness that is *entirely*

immune to functional analysis, then it must be a content which has no causal effects, which is causally ineffective (disregarding non-functional causal influences). This must be a curious content indeed, since anything which is entirely causally ineffective must also be entirely unremarkable in the literal sense; it will also be literally unmemorable. Indeed it is a strange thought that philosophers can coherently talk about such content at all. The use of the concept of qualia to argue for such ineffable objects appears to be a generally confused one: such `qualophiles' want to argue both that conscious content is effective (we drink the water because we are conscious of thirst) and that the same conscious content is ineffective (the `precise' feel of thirst cannot be expressed or analyzed into any functional relations). It is in this refinedly mysterious sense of consciousness that Dennett is compelled to announce that we are none of us conscious. I automatically assent.

5. Conclusion

5.1 I hope to have given some impression of the range of topics, without pretending to have surveyed them. As I have made clear, there is much in this book that is disputable. And Dennett is at times aggravatingly smug and confident about the merits of his arguments (comparing his `revelations' about consciousness to a magician's revealing the operation of stage tricks, for example; p. 434). All in all Dennett's book is annoying, frustrating, insightful, provocative and above all annoying. Unfortunately---in this age of academic overproduction---I must conclude that for now *Consciousness Explained* is unavoidable reading for those who intend to think seriously about the problems of consciousness.

Notes

<1> To be sure, the context of the quote is packed with qualifiers and a footnote complains about my quoting him out of context. But honestly how could I resist?}

<2> Operationalism claims that the content of our concepts is exhausted by the operations we use to measure or judge their application. This idea presupposes that we cannot find new means to measure or judge the presence of previously defined concepts, and is refuted, for example, by the history of the concept of temperature.

<3> In the replies to commentaries section of Dennett and Kinsbourne (1992) my point here is evidently acknowledged. Presumably, Dennett's target is those who insist on asking *precisely* when or *exactly* where consciousness begins and ends---and trying to mean by that a degree of precision that is unobtainable in principle. But while such requests are misleading, and puzzles based upon them pointless, it is also misleading in another way to insist flatly that there is no fact of the matter as to where and when consciousness occurs. (Incidently, readers of *Consciousness Explained* will find Dennett and Kinsbourne (1992) of interest---especially, perhaps, the numerous peer commentaries on Dennett and Kinsbourne's description of the Multiple Drafts model and their replies thereto.)

References

Churchland, P.M. (1981). Eliminative materialism and propositional attitudes. *Journal of Philosophy, 78,* 67-90.

Dennett, D. (1969). *Content and consciousness*. London: Routledge & Kegan Paul.

Dennett, D. (1979). On the absence of phenomenology. In D. Gustafson & B. Tapscott (Eds.) *Body, mind and method: Essays in honor of Virgil Aldrich*. Dordrecht Reidel.

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.

Kolers, P.A., & von Gruenau, M. (1979). Shape and color in apparent motion. *Vision Research, 16,* 329-335.

Korb, K. (1991). Searle's AI program. *Journal of Experimental and Theoretical AI, 3,* 283-296.

Lewis, D. (1980). Mad pain and Martian pain. In N. Block (Ed.) *Readings in philosophy of psychology: Vol. 1* (pp. 216-222).

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 49,* 329-359.

Rosenthal, D. (1990). *A theory of consciousness* (ZIF Report No. 40). Zentrum fuer Interdisziplinaere Forschung, Bielefeld, Germany.

Selfridge, O. (1959). Pandemonium: a paradigm for learning. In *Symposium on the Mechanization of Thought Processes*. London: HM Stationery Office.