Philosophy and Phenomenological Research
Vol. LIII, No. 4, December 1993

Multiple Drafts and Higher-Order Thoughts

DAVID M. ROSENTHAL
City University of New York, Graduate Center

One strategy for explaining consciousness is to locate it relative to its causes and effects. Common sense seems to tell us what the causes and effects of consciousness are; so perhaps we can identify consciousness as being whatever it is that occurs in between the two. Though superficially tempting, this idea heightens the air of mystery surrounding consciousness. As far as we can tell, the physiological causes of consciousness lead directly to its physiological effects. So how could consciousness lie anywhere along that causal path unless, as Dennett evocatively puts it, "a miracle occurs"?[1]

The idea that consciousness occurs at the unique interface of certain causes and effects underlies what Dennett calls the Cartesian Theater model of mind. I agree with Dennett that this model, though seldom acknowledged, tacitly underlies much that's mistaken in current thinking about mind and consciousness; his sustained, effective effort to expose and demolish that myth is an important contribution to our understanding of consciousness. I'm convinced, moreover, that a correct explanation of consciousness will share much with the alternative theory Dennett puts forth, the Multiple Drafts model (MDM). Nonetheless, I think the explanatory benefits of the MDM may be available with a model that's in one respect somewhat weaker.

Perhaps the most important virtue of the MDM is its ability to accommodate the apparent temporal anomalies Dennett catalogues. In color phi a green flash occurs in the left of our visual field followed by a red flash on the right, but we seem to see a single spot that moves and changes color. Why don't we first consciously sense the initial stationary flash? Dennett believes the MDM has the answer.

According to the MDM, consciousness is continuously revised, much as a text changes through successive drafts. Some features of a text undergoing revision persist through many drafts; others may be so transitory as to es-

[1] Daniel C. Dennett, *Consciousness Explained*, Boston: Little, Brown and Company, 1991, p. 38. Page references throughout are to this book.

cape notice altogether. Similarly with consciousness. If we try to determine what conscious experience a subject has, eliciting a reaction at different times may well result in different reports. At successive moments we might "precipitat[e] different narratives...: [different] versions of a portion of 'the stream of consciousness'" (135). Nor is there any privileged moment at which a report would reveal the true nature of the subject's conscious experience. Even from a first-person viewpoint, introspective impressions of our own experiences sometimes vary through time. Reporting is like publishing a text. But even publication fixes a text only relative to a social context; post-publication revision can and does occur.

These points help explain color phi. If the initial stationary flash enters the subject's consciousness at all, its presence is so transitory that it escapes notice altogether. Perhaps at some early moment we could elicit a report of the stationary flash; but if we don't, it's as though no conscious experience of the stationary flash ever occurred.

This much is fairly straightforward. In many cases we simply won't know, from either a first- or third-person point of view, whether a particular stimulus makes it to consciousness. But Dennett's MDM goes one step farther, and "makes 'writing it down' in memory criterial for consciousness.... There is no reality of consciousness independent of the effects of various vehicles of content on subsequent action (and hence, of course, on memory)" (132). Accordingly, Dennett rejects both the "Stalinesque" view, on which the initial flash is edited out prior to consciousness, and the opposing "Orwellian" claim that it reaches consciousness but is immediately forgotten. He holds that when no early reaction is elicited, there's no fact of the matter about whether the initial stimulus makes it to consciousness. "[T]here are no fixed facts about the stream of consciousness independent of particular probes" (138; cf. 275).

Central to the MDM is the idea that it's not fixed from one moment to the next what our conscious experiences are. A satisfactory explanation of consciousness will very likely have to accommodate such revisability. But it's less clear that we must also adopt Dennett's "*first-person operationalism*" (132): his denial that there's any fact of the matter about whether certain stimuli reach consciousness, and about what our conscious experiences are at any particular moment.

We can explore this question by seeing whether the explanatory virtues of the MDM remain when we subtract first-person operationalism, that is, when we adopt a theory that affirms the revisability just described but denies first-person operationalism. One such theory is the higher-order-thought explanation of consciousness that I've put forth elsewhere, and which Dennett discusses in chapter 10. When a mental state is conscious, one is conscious of being in that state. On the higher-order-thought hypothesis, one is conscious of a conscious state by virtue of having a roughly contempo-

raneous thought to the effect that one is in that state. Because that thought is about another mental state, it's convenient to call it a higher-order thought (HOT). We sometimes infer that we're in some mental state even when that state isn't conscious; so we must stipulate that the HOT is independent of any inference of which we're aware. Normally, we're not conscious of the HOTs this theory posits, but that's to be expected; a HOT isn't conscious unless one has a third-order thought about it, which seldom happens.[2]

The HOT hypothesis accommodates the apparent temporal anomalies no less well than the MDM. The initial flash in color phi presumably causes a sensory state of a stationary colored spot. But it may well be that no HOT about that sensory state occurs; so the sensory state never becomes conscious. After the second flash occurs on the right, a new sensory state occurs of a moving spot that changes color. A HOT about this second sensory state then occurs, giving one the conscious experience of a moving, changing spot.

There's also, however, a second possibility. A HOT might occur about the initial sensory state of a stationary red flash, but be replaced so fast by the other HOT about the sensory state of a moving spot that the first HOT doesn't register mentally. It doesn't last long enough to affect our memory of what we saw, and it has no other noticeable effects. In particular, there's no time to report that one is in the first sensory state, that is, to *express* one's HOT about that state. The first sensory state is conscious, but so briefly as to make no mental difference.

These two possibilities instantiate the Stalinesque and Orwellian models, respectively. Because the two scenarios are introspectively indistinguishable, Dennett is right that introspection can't determine which has occurred. He's also right that nonverbal behavioral reactions won't help, since, as he argues, any such reaction could equally well result from a state that's conscious or a state that's not (124).

It doesn't follow, however, that there's no way to tell which model is operative in a particular case. There may be theoretical reasons, beyond what pertains to small time scales, for distinguishing between a stimulus's not becoming conscious and its becoming conscious but not seeming to. Moreover, the HOT hypothesis could provide such reasons, since the Orwellian

---

2 See David M. Rosenthal, "Thinking that One Thinks," in *Consciousness: A Mind and Language Reader*, ed. Glyn W. Humphreys and Martin Davies (Oxford: Basil Blackwell, 1993), pp. 197–223; "The Independence of Consciousness and Sensory Quality," in *Consciousness: Philosophical Issues, 1, 1991*, ed. Enrique Villanueva (Atascadero, California: Ridgeview Publishing Company, 1991), pp. 15–36; "Two Concepts of Consciousness," *Philosophical Studies* 49, 3 (May 1986), pp. 329–59; and "Why Are Verbally Expressed Thoughts Conscious?" and "A Theory of Consciousness," Reports 32 and 40, Center for Interdisciplinary Research (ZiF), Research Group on Mind and Brain, University of Bielefeld.

I use 'thought' here as a term of art for any episodic intentional state with an assertoric mental attitude.

model posits an initial HOT that doesn't occur on the Stalinesque model. We don't now know, of course, which model explains color phi or the other temporal anomalies. Perhaps some anomalies are Orwellian and others Stalinesque; perhaps some have both Stalinesque and Orwellian instances. Still, explaining these phenomena doesn't require us to deny that there's a fact of the matter about which occurs.

Arguably, the HOT hypothesis explains the many other mental phenomena Dennett considers at least as well as the MDM. And first-person operationalism aside, the MDM and HOT hypothesis agree about most theoretical issues. On the Cartesian Theater model, consciousness is a unifying factor, since it brings one's conscious states together in a single mental place. The MDM insists instead that consciousness is a distributed phenomenon; there's no one place where consciousness occurs. So there's no unique Central Meaner and no unique Author of Record (228). The HOT hypothesis agrees. Distinct mental states are conscious because of different HOTs, which presumably occur in different locations.[3] Dennett explains the apparent unity of consciousness as due to the way our mental states are linked in the narratives we construct. The HOT hypothesis takes a similar line; unity results from HOTs that subsume groups of mental states, and from occasional third-order thoughts that in turn connect several HOTs. Indeed, Dennett's narratives are just the linguistic expressions of these HOTs.

On the MDM, a mental state is conscious if it leaves significant traces in memory, and has other substantial mental and behavioral effects. But not all a mental state's effects are relevant to whether it's conscious. As Dennett notes, conscious and nonconscious mental states can have the same effects on nonverbal behavior. Similarly, most of the mental traces conscious states leave could equally have been left by nonconscious mental states. This is true even of effects on memory, which Dennett counts as criterial for consciousness. One sometimes sees something without being at all conscious of seeing it, even though one later recalls having seen it—perhaps to one's surprise. In these cases, nonconscious perceiving has a significant, lasting effect on memory.

A mental state can have many mental effects without becoming conscious, but not if it causes a HOT. Having a thought about something is one way of being conscious of it. So if one comes noninferentially to have a thought that one is in a particular mental state, that state becomes conscious. The HOT hypothesis thus focuses more closely than the MDM on the mental effects mental states must have to be conscious.

Dennett emphasizes that it's not always clear, even from a first-person point of view, whether one is conscious of something, illustrating this with

---

3 So heterophenomenology needn't presuppose a single Author of Record, as Dennett claims (228).

the game of "hide the thimble" (336). This poses a problem. The sensory states near the center of one's visual field are normally conscious; so if one is looking at the hidden thimble, how can one fail to see it consciously? Moreover, the difficulty we have in describing this kind of case from a first-person point of view seems to support Dennett's claim that, independent of particular probes, there's no fact of the matter about what conscious experiences we have.

We're seldom if ever conscious of all the detail that's represented in our sensory states, even sensory states at the center of our visual field. And the amount of detail we're conscious of often changes. When that happens, moreover, it needn't be the sensory state that changes, but only the way we're conscious of that state. The HOT hypothesis explains these things. HOTs represent sensory states in greater or lesser detail. So, a HOT might represent one's sensory state as being just of a bookcase with lots of things on it. But the HOT might instead represent the sensory state in greater detail, as including a thimble. In the first case one is conscious of seeing the bookcase but not the thimble; in the second case one's conscious of seeing both. All this is independent of first-person operationalism.

As Dennett notes, training in such things as piano tuning and wine tasting can change what conscious experiences one has. But this training by itself can hardly produce new kinds of sensory states; rather, we acquire new discriminatory concepts for our experiences, and the resulting HOTs provide finer detail in the way we're conscious of those experiences. Dennett speculates that if blindsight patients became self-cuing, the states in their "blind" hemifield would become conscious. The HOT hypothesis suggests why; successfully guessing when to guess about one's blind hemifield would lead to having HOTs that one's sensing something there.

Given the low resolution of parafoveal vision, Dennett convincingly argues that we see a wide area of wallpaper as being all Marilyns by representing a few foveal Marilyns plus a judgment that the parafoveal shapes are "more of the same" (355). The HOT hypothesis suggests how this might happen. The relevant sensory state represents foveal shapes as Marilyns and peripheral shapes as largely indistinct shapes. One's HOT, then, cleans things up being a thought that one is in a sensory state in which foveal and peripheral shapes are all Marilyns. And more generally, the HOT hypothesis fits well with Dennett's view that so-called "filling in" takes place not by the brain's manufacturing the requisite sensory states, but by its forming the requisite judgments.

Dennett believes that the heterophenomenological method undercuts the opposition between first- and third-person viewpoints. Again, the HOT hypothesis helps explain how. Heterophenomenological reports express HOTs about the person's conscious states. Indeed, since we're conscious of our con-

scious states by virtue of HOTs, heterophenomenological reports are the most revealing expression of a subject's consciousness of those states.[4]

Dennett propounds first-person operationalism in part because its denial "creates the bizarre category of the objectively subjective—the way things actually, objectively seem to you even if they don't seem to seem that way to you" (132). Moreover, he argues that the acceptance of this distinction by the HOT hypothesis is one reason to reject that hypothesis (316).

There are, however, good reasons to sustain a distinction between seeming and seeming to seem. Being in a sensory state defines, in one respect, how things seem to me. That's because of connections that state has to other aspects of my mental life, even if the state fails to be conscious. But when the sensory state isn't conscious, I'm not aware of being in it; so it doesn't then seem to me that I'm in it. That's the second level of seeming.[5] Because Dennett holds that leaving significant traces and having mental effects is what it is for a state to be conscious, he might deny that a nonconscious sensory state can define how things seem to one. But as already argued, mental states can have many mental effects without being conscious.

Dennett might urge that ascribing representational properties to sensory states incurs all the problems that face qualia and the mental "pigment" he argues against (346). But arguably mental pigment and qualia are problematic only because we conceive of them as being intrinsically conscious.[6] On the HOT hypothesis, the representational properties of sensory states need not occur consciously.

Dennett notes that distinguishing a mental state from the corresponding HOT makes room for an unexpected kind of error; my HOT might misrepresent what mental state I'm in (317). And if there's a third-order thought, it might in turn be wrong about what my second-order thought is.[7] Dennett regards these proliferating levels of possible error as another reason to reject

---

4 Accordingly, the MDM and HOT hypothesis agree about zombies (and zimboes [310–11]). On the HOT hypothesis, having HOTs is sufficient for a state to be conscious; so nothing could have all our intentional states but lack conscious states. Similarly, Dennett denies that anything "in principle...indistinguishable from a conscious person" could lack conscious states (405; cf. 282).

5 Dennett seems to say as much: the "onsets [of content-fixations in the brain] do *not* mark the onset of consciousness of their content" (113; emphasis Dennett's). Not distinguishing these two levels, moreover, risks representing consciousness as an intrinsic property of mental states, which accords poorly with Dennett's idea that consciousness is a distributed phenomenon. And if being conscious were an intrinsic property, it's unlikely we could explain what it is for a state to be conscious without appealing to the very consciousness we were trying to explain.

6 See especially Rosenthal, "The Independence of Consciousness and Sensory Quality."

7 Dennett takes the HOT hypothesis to posit not just HOTs, but also higher-order beliefs distinct from those HOTs, apparently because he assumes that reports of conscious states express higher-order beliefs (307; cf. 317). He concludes that error might also occur between HOT and belief. But reports of conscious states directly express HOTs; so the HOT hypothesis doesn't posit higher-order beliefs at all.

the HOT hypothesis, since he thinks it's idle to distinguish among mistakes at these different levels.

We're never aware of all the introspectible features of our conscious states; consider the hidden thimble. So the way we're conscious of our conscious states is never wholly accurate. Why, then, shouldn't the way we're conscious of those states sometimes represent them as having features they lack? There's evidence from clinical contexts and from social psychology[8] that this sometimes occurs. Even introspective impressions may occasionally be erroneous because one's third-order thought misrepresents the content of a second-order thought. The two kinds of error can be distinct even if they're not introspectively distinguishable.[9]

Because HOTs can misrepresent mental states, we'd need independent evidence to tell which of two HOTs is more accurate. And since heterophenomenological reports express HOTs, the HOT hypothesis holds, with the MDM, that there's no privileged moment at which those reports determine the true nature of a subject's conscious experience. The possibility of erroneous HOTs doesn't imply first-person operationalism. Facts about what mental states we're in differ from facts about how we're conscious of them, and HOTs pertain only to the latter.

Dennett argues that our choice of words can influence the content of our thoughts (247), partly because we often discover what we think only as we say it (245). He concludes that HOTs don't always underlie heterophenomenological reports (315). But we may discover what we think as we say it not because our words affected what the thought was, but because our thought hadn't been conscious before we spoke. And even if reports of mental states sometimes do affect the content of our HOTs, each report of a mental state still expresses a HOT in virtue of which that state is conscious.

Dennett sees as artificial the ordinary way we carve consciousness and mind into discrete mental states. Worse, he thinks it results in "postulating differences that are systemically undiscoverable." It's at bottom because the HOT hypothesis and the Stalinesque and Orwellian models proceed in this way that he rejects both the HOT hypothesis and the distinction between Stalinesque and Orwellian explanations (319). But Dennett himself writes of such things as "events of content-fixation" (365), "information-bearing events" (459), "content-discriminations" (113), and narrative fragments getting "precipitated" (136). Moreover, these "content-fixations...are

---

8 Richard E. Nisbett and Timothy DeCamp Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84, 3 (May 1977): 231–59.

9 Because Dennett sees the HOT hypothesis as reflecting folk psychology, he regards both as committed to these distinct levels of possible error. It's unclear that these distinct levels of error are unequivocally part of our folk-psychological picture. But in any event, the HOT hypothesis is a theoretical proposal about what it is for mental states to be conscious, and so isn't constrained by folk-psychological conceptions.

[each] precisely locatable in both space and time" (113). It's unclear how he thinks these events differ from the mental states presupposed by the HOT hypothesis and the Stalinesque and Orwellian models. Without additional support for first-person operationalism, then, we may hope to explain consciousness by a version of the MDM that lacks those operationalist consequences.