BEHAVIORAL AND BRAIN SCIENCES (1992) 15, 183–247
Printed in the United States of America

Time and the observer: The where and when of consciousness in the brain

Daniel C. Dennett[a] and Marcel Kinsbourne[b]
[a]Center for Cognitive Studies, Tufts University, Medford, MA 02155;
[b]Behavioral Neurology Unit, Sargent College, Boston University, Boston, MA 02215
Electronic mail: ddennett@pearl.tufts.edu

Abstract: We compare the way two models of consciousness treat subjective timing. According to the standard "Cartesian Theater" model, there is a place in the brain where "it all comes together," and the discriminations in all modalities are somehow put into registration and "presented" for subjective judgment. The timing of the events in this theater determines subjective order. According to the alternative "Multiple Drafts" model, discriminations are distributed in both space and time in the brain. These events do have temporal properties, but those properties do not determine subjective order because there is no single, definitive "stream of consciousness," only a parallel stream of conflicting and continuously revised contents. Four puzzling phenomena that resist explanation by the Cartesian model are analyzed: (1) a gradual apparent motion phenomenon involving abrupt color change (Kolers & von Grünau 1976), (2) an illusion of an evenly spaced series of "hops" produced by two or more widely spaced series of taps delivered to the skin (Geldard & Sherrick's "cutaneous rabbit" [1972]), (3) backwards referral in time, and (4) subjective delay of consciousness of intention (both reported in this journal by Libet 1985a; 1987; 1989a). The unexamined assumptions that have always made the Cartesian Theater so attractive are exposed and dismantled. The Multiple Drafts model provides a better account of the puzzling phenomena, avoiding the scientific and metaphysical extravagances of the Cartesian Theater: The temporal order of subjective events is a product of the brain's interpretational processes, not a direct reflection of events making up those processes.

Keywords: consciousness; discrimination; illusion; localization; memory; mental timing; perception; subjective experience

> I'm really not sure if others fail to perceive me or if, one fraction of a second after my face interferes with their horizon, a millionth of a second after they have cast their gaze on me, they already begin to wash me from their memory: forgotten before arriving at the scant, sad archangel of a remembrance.
> Ariel Dorfman, *Mascara*, 1988

When scientific advances contradict "common sense" intuitions, the familiar ideas often linger on, not just outliving their usefulness but even confusing the scientists whose discoveries ought to have overthrown them. Diagnosed here is a ubiquitous error of thinking that arises from just such a misplaced allegiance to familiar images, illustrated with examples drawn from recent work in psychology and neuroscience. Although this is a "theoretical" paper, it is addressed especially to those who think, mistakenly, that they have no theories and no need for theories. We show how uncontroversial facts about the spatial and temporal properties of information-bearing events in the brain require us to abandon a family of entrenched intuitions about "the stream of consciousness" and its relation to events occurring in the brain.

In Section 1, we introduce two models of consciousness, the standard Cartesian Theater and our alternative, the Multiple Drafts model, briefly describing four phenomena of temporal interpretation that raise problems for the standard model. Two of these, drawn from the research of Libet, have been extensively debated on methodological grounds, but concealed in the controversy surrounding them are the mistaken assumptions we expose. In Section 2, we diagnose these intuitive but erroneous ideas and exhibit their power to create confusion in relatively simple contexts. We demonstrate the superiority of the Multiple Drafts model of consciousness by showing how it avoids the insoluble problems faced by versions of the Cartesian Theater. In Section 3, we show how covert allegiance to the Cartesian Theater has misled interpreters of Libet's phenomena and how the Multiple Drafts model avoids these confusions.

1. Two models of consciousness

1.1. Cartesian materialism: Is there a "central observer" in the brain? Wherever there is a conscious mind, there is a *point of view*. A conscious mind is an observer who takes in the information that is available at a particular (roughly) continuous sequence of times and places in the universe. A mind is thus a *locus of subjectivity*, a thing it is like something to be (Farrell 1950; Nagel 1974). What it is like to be that thing is partly determined by what is available to be observed or experienced along the trajectory through space-time of that moving point of view, which for most practical purposes is just that: *a point*. For instance, the startling dissociation of the sound and appearance of distant fireworks is explained by the different

 0140-525X/92 $5.00+.00

transmission speeds of sound and light, arriving *at the observer* (at that point) at different times, even though they left the source simultaneously. But if we ask where precisely in the brain that point of view is located, the simple assumptions that work so well on larger scales of space and time break down. It is now quite clear that there is no single point in the brain where all information funnels in, and this fact has some far from obvious consequences.

Light travels much faster than sound, as the fireworks example reminds us, but it takes longer for the brain to process visual stimuli than to process auditory stimuli. As Pöppel (1985/1988) has pointed out, thanks to these counterbalancing differences, the "horizon of simultaneity" is *about* 10 meters: Light and sound that leave the same point about 10 meters from the observer's sense organs produce neural responses that are "centrally available" at the same time. Can we make this figure more precise? There is a problem. The problem is not just measuring the distances from the external event to the sense organs, or the transmission speeds in the various media, or allowing for individual differences. The more fundamental problem is deciding what to count as the "finish line" in the brain. Pöppel obtained his result by comparing behavioral measures: mean reaction times (button-pushing) to auditory and visual stimuli. The difference ranges between 30 and 40 msec, the time it takes sound to travel approximately 10 meters (the time it takes light to travel 10 meters is only infinitesimally different from zero). Pöppel used a peripheral finish line – external behavior – but our natural intuition is that the *experience* of the light and sound happens *between* the time the vibrations strike our sense organs and the time we manage to push the button to signal that experience. And it happens somewhere *centrally*, somewhere in the brain on the excited paths between the sense organ and muscles that move the finger. It seems that if we could say exactly *where* the experience happened, we could infer exactly *when* it happened. And vice versa: If we could say exactly when it happened, we could infer where in the brain conscious experience was located.

This picture of how conscious experience must sit in the brain is a natural extrapolation of the familiar and undeniable fact that *for macroscopic time intervals*, we can indeed order events into the categories "not yet observed" and "already observed" by locating the observer and plotting the motions of the vehicles of information relative to that point. But when we aspire to extend this method to explain phenomena involving very short intervals, we encounter a *logical* difficulty: If the "point" of view of the observer is spread over a rather large volume in the observer's brain, the observer's own subjective sense of sequence and simultaneity *must* be determined by something other than a unique "order of arrival" because order of arrival is incompletely defined until we specify the relevant destination. If A beats B to one finish line but B beats A to another, which result fixes subjective sequence in consciousness (cf. Minsky 1985, p. 61)? Which point or points of "central availability" would "count" as a determiner of *experienced* order, and why?

Consider the time course of normal visual information processing. Visual stimuli evoke trains of events in the cortex that gradually yield content of greater and greater specificity. At different times and different places, various "decisions" or "judgments" are made: More literally, parts of the brain are caused to go into states that differentially respond to different features, for example, first mere onset of stimulus, then shape, later color (in a different pathway), motion, and eventually object recognition. It is tempting to suppose that there must be some place in the brain where "it all comes together" in a multimodal representation or display that is *definitive* of the content of conscious experience in at least this sense: The temporal properties of the events that occur in that particular locus of representation determine the temporal properties – of sequence, simultaneity, and real-time onset, for instance – of the subjective "stream of consciousness." This is the error of thinking we intend to expose. Where does it all "come together?" The answer, we propose, is nowhere. Some of the contentful states distributed around in the brain soon die out, leaving no traces. Others do leave traces, on subsequent verbal reports of experience and memory, on "semantic readiness" and other varieties of perceptual set, on emotional state, behavioral proclivities, and so forth. Some of these effects – for instance, influences on subsequent verbal reports – are at least symptomatic of consciousness. But there is no one place in the brain through which all these causal trains must pass to deposit their contents "in consciousness" (see also Damasio 1989a).

The brain must be able to "bind" or "correlate" and "compare" various separately discriminated contents, but the processes that accomplish these unifications are themselves distributed, not gathered at some central decision point, and as a result, the "point of view of the observer" is spatially smeared. If brains computed at near the speed of light, as computers do, this spatial smear would be negligible. But given the relatively slow transmission and computation speeds of neurons, the spatial distribution of processes creates significant temporal smear – ranging, as we shall see, up to several hundred milliseconds – within which range the normal commonsense assumptions about timing and arrival at the observer need to be replaced. For many tasks, the human capacity to make conscious discriminations of temporal order drops to chance when the difference in onset is on the order of 50 msec (depending on stimulus conditions), but this variable threshold is the result of complex interactions, not a basic limit on the brain's capacity to make the specialized order judgments required in the interpretation and coordination of perceptual and motor phenomena. We need other principles to explain the ways *subjective temporal order* is composed, especially in cases in which the brain must cope with rapid sequences occurring at the limits of its powers of temporal resolution. As usual, the performance of the brain when put under strain provides valuable clues about its general modes of operation.

Descartes, early (1664) to think seriously about what must happen inside the body of the observer, elaborated an idea that is superficially so natural and appealing that it has permeated our thinking about consciousness ever since and permitted us to defer considering the perplexities – until now. Descartes decided that the brain *did* have a center: the pineal gland, which served as the gateway to the conscious mind. This was the only organ in the brain that was in the midline, rather than paired, with left and right versions. The pineal looked different, and because its function was then quite inscrutable (and still

is), Descartes posited a role for it: For a person to be conscious of something, traffic from the senses had to arrive at this station, where it thereupon caused a special – indeed magical – transaction to occur between the person's material brain and immaterial mind. When the conscious mind then decided on a course of bodily action, it sent a message back "down" to the body via the pineal gland. The pineal gland, then, is like a theater in which information is displayed for perusal by the mind.

Descartes' vision of the pineal's role as the turnstile of consciousness (we might call it the Cartesian bottleneck) is hopelessly wrong. The problems that face Descartes' interactionistic dualism, with its systematically inexplicable traffic between the realm of the material and the postulated realm of the immaterial, were already well appreciated in Descartes' own day, and centuries of reconsideration have only hardened the verdict: The idea of the Ghost in the Machine, as Ryle (1949) aptly pilloried it, is a nonsolution to the problems of mind. But whereas materialism of one sort or another is now a received opinion approaching unanimity,[1] even the most sophisticated materialists today often forget that once Descartes' ghostly *res cogitans* is discarded, there is no longer a role for a centralized gateway, or indeed for any *functional* center to the brain. The brain itself is Headquarters, the place where the ultimate observer is, but it is a mistake to believe that the brain has any deeper headquarters, any inner sanctum, arrival at which is the necessary or sufficient condition for conscious experience.

Let us call the idea of such a centered locus in the brain *Cartesian materialism*, because it is the view one arrives at when one discards Descartes' dualism but fails to discard the associated imagery of a central (but material) theater where "it all comes together." Once made explicit, it is obvious that this is a bad idea, not only because, as a matter of empirical fact, nothing in the functional neuroanatomy of the brain suggests such a general meeting place, but also because positing such a center would apparently be the first step in an infinite regress of too-powerful homunculi. If all the tasks Descartes assigned to the immaterial mind have to be taken over by a "conscious" *sub*system, its own activity will either be systematically mysterious or decomposed into the activity of further subsystems that begin to duplicate the tasks of the "nonconscious" parts of the whole brain. Whether or not anyone explicitly endorses Cartesian materialism, some ubiquitous assumptions of current theorizing presuppose this dubious view. We show that the persuasive imagery of the Cartesian Theater, in its materialistic form, keeps reasserting itself, in diverse guises, and for a variety of ostensibly compelling reasons. Thinking in its terms is not an innocuous shortcut; it is a bad habit. One of its most seductive implications is the assumption that a distinction can *always* be drawn between "not yet observed" and "already observed." But, as we have just argued, this distinction *cannot* be drawn once we descend to the scale that places us within the boundaries of the spatiotemporal volume in which the various discriminations are accomplished. Inside this expanded "point of view," spatial and temporal distinctions lose the meanings they have in broader contexts.

The crucial features of the Cartesian Theater model can best be seen by contrasting it with the alternative we propose, the Multiple Drafts model:

All perceptual operations, and indeed all operations of thought and action, are accomplished by multitrack processes of interpretation and elaboration that occur over hundreds of milliseconds, during which time various additions, incorporations, emendations, and overwritings of content can occur, in various orders. Feature-detections or discriminations *have to be made only once*. That is, once a localized, specialized "observation" has been made, the information content thus fixed does not have to be sent somewhere else to be *re*discriminated by some "master" discriminator. In other words, it does not lead to a re-*presentation* of the already discriminated feature for the benefit of the audience in the Cartesian Theater. How a localized discrimination contributes to, and what effect it has on the prevailing brain state (and thus awareness) can change from moment to moment, depending on what else is going on in the brain. Drafts of experience can be revised at a great rate, and no one is more correct than another. Each reflects the situation at the time it is generated. These spatially and temporally distributed content-fixations are themselves precisely locatable in both space and time, but their onsets do *not* mark the onset of awareness of their content. It is always an open question whether any particular content thus discriminated will eventually appear as an element in conscious experience. These distributed content-discriminations yield, over the course of time, something *rather like* a narrative stream or sequence, subject to continual editing by many processes distributed around in the brain, and continuing indefinitely into the future (cf. Calvin's [1990] model of consciousness as "scenario-spinning"). This stream of contents is only rather like a narrative because of its multiplicity; at any 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 intervals produces different effects, elicits different narrative accounts from the subject. If one delays the probe too long (overnight, say) the result is apt to be no narrative left at all – or else a narrative that has been digested or "rationally reconstructed" to the point that it has minimal integrity. If one probes "too early," one may gather data on how early a particular discrimination is achieved in the stream, but at the cost of disrupting the normal progression of the stream. Most important, the Multiple Drafts model avoids the tempting mistake of supposing that there must be a single narrative (the "final" or "published" draft) that is canonical – that represents the *actual* stream of consciousness of the subject, whether or not the experimenter (or even the subject) can gain access to it.

The main points at which this model disagrees with the competing tacit model of the Cartesian Theater, may be summarized:

1. Localized discriminations are *not* precursors of re-*presentations* of the discriminated content for consideration by a more central discriminator.

2. The objective temporal properties of discriminatory states may be determined, but they do *not* determine temporal properties of subjective experience.

3. The "stream of consciousness" is *not* a single, definitive narrative. It is a parallel stream of conflicting and continuously revised contents, no one narrative thread of which can be singled out as canonical – as the true version of conscious experience.

The different implications of these two models will be exhibited by considering several puzzling phenomena that seem at first to indicate that the mind "plays tricks with time." (Other implications of the Multiple Drafts model are examined at length in Dennett 1991b.)

1.2. Some "temporal anomalies" of consciousness. Under various conditions people report experiences in which the temporal ordering of the elements in their consciousness, or the temporal relation of those elements to concurrent activity in their brains, seems to be anomalous or even paradoxical. Some theorists (Libet 1982; 1985a; Popper & Eccles 1977) have argued that these temporal anomalies are proof of the existence of an immaterial mind that interacts with the brain in physically inexplicable fashion. Others (Goodman 1978; Libet 1985b), although eschewing any commitment to dualism, have offered interpretations of the phenomena that seem to defy the accepted temporal sequence of cause and effect. Most recently, another theorist (Penrose 1989 – see also multiple book review in *BBS* 13 (4) 1990) has suggested that a materialistic explanation of these phenomena would require a revolution in fundamental physics. These radical views have been vigorously criticized, but the criticisms have overlooked the possibility that the appearance of anomaly in these cases results from conceptual errors that are so deeply anchored in everyday thinking that even many of the critics have fallen into the same traps. We agree with Libet and others that these temporal anomalies are significant, but we hold a different opinion about what they signify.

We focus on four examples, summarized below. Two, drawn from the work of Libet, have received the most attention and provoked the most radical speculation, but because technical criticisms of his experiments and their interpretation raise doubts about the existence of the phenomena he claims to have discovered, we begin with a discussion of two simpler phenomena whose existence has not been questioned but whose interpretation raises the same fundamental problems. We use these simpler cases to illustrate the superiority of the Multiple Drafts model to the traditional Cartesian Theater model, and then apply the conclusions drawn in the more complicated setting of the controversies surrounding Libet's work. Our argument is that even if Libet's phenomena were not known to exist, theory can readily account for the possibility of phenomena of this pseudo-anomalous sort, and even predict them.

A. Color phi. Many experiments have demonstrated the existence of apparent motion, or the phi phenomenon (Kolers & von Grünau 1976; see also Kolers 1972; van der Waals & Roelofs 1930; and the discussion in Goodman 1978). If two or more small spots separated by as much as 4 degrees of visual angle are briefly lit in rapid succession, a single spot will seem to move. This is the basis of our experience of motion in motion pictures and television. First studied systematically by Wertheimer (1912; for a historical account, see Kolers 1972; Sarris 1989), phi has been subjected to many variations; one of the most striking is reported in Kolers and von Grünau (1976). The philosopher Nelson Goodman had asked Kolers whether the phi phenomenon would persist if the two illuminated spots were different in color, and if so, what would happen to the color of "the" spot as "it" moved? Would the illusion of motion disappear, to be replaced by two separately flashing spots? Would the illusory "moving" spot gradually change from one color to another, tracing a trajectory around the color wheel? The answer, when Kolers and von Grünau performed the experiments, was striking: The spot seems to begin moving and then to change color abruptly *in the middle of its illusory passage* toward the second location. Goodman wondered: "How are we able . . . to fill in the spot at the intervening place-times along a path running from the first to the second flash *before that second flash occurs*?" (1978, p. 73; the same question can be raised about any phi, but the color-switch in midpassage vividly brings out the problem.) Unless there is precognition, the illusory content cannot be created until *after* some identification of the second spot occurs in the brain. But if this identification of the second spot is already "in conscious experience" would it not be too late to interpose the illusory color-switching-while-moving scene between the conscious experience of spot 1 and the conscious experience of spot 2? How does the brain accomplish this sleight-of-hand? Van der Waals and Roelofs (1930) proposed that the intervening motion is produced retrospectively, built only after the second flash occurs, and "projected backwards in time" (Goodman 1978, p. 74), a form of words reminiscent of Libet's "backwards referral in time." But what does it mean that this experienced motion is "projected backwards in time"?

B. The cutaneous "rabbit." The subject's arm rests cushioned on a table, and mechanical square-wave tappers are placed at two or three locations along the arm, up to a foot apart (Geldard & Sherrick 1972; see also Geldard 1977; Geldard & Sherrick 1983; 1986). A series of rhythmical taps is delivered, for example, 5 at the wrist followed by 2 near the elbow and then 3 more on the upper arm. These taps are delivered with interstimulus intervals of between 50 and 200 msec. So a train of taps might last less than a second, or as long as two or three seconds. The astonishing effect is that the taps seem to the subjects to travel in regular sequence over equidistant points up the arm – as if a little animal were hopping along the arm. Now *how did the brain know* that after the 5 taps on the wrist there were going to be some taps near the elbow? The experienced "departure" of the taps from the wrist begins with the second one, yet in catch trials in which the later elbow taps are never delivered, all five wrist taps are felt at the wrist in the expected manner. The brain obviously cannot "know" about a tap at the elbow until after it happens. Perhaps, one might speculate, the brain delays the conscious experience until after all the taps have been "received" and then, somewhere upstream of the seat of consciousness (whatever that is), *revises* the data to fit a theory of motion, and sends the edited version on to consciousness. But would the brain always delay response to one tap in case more came? If not, how does it "know" when to delay?

C. "Referral backwards in time." Since Penfield and Jasper (1954) it has been known that direct electrical stimulation of locations on the somatosensory cortex can induce sensations on corresponding parts of the body. For instance, stimulation of a point on the left somatosensory

cortex can produce the sensation of a brief tingle in the subject's right hand. Libet compared the time course of such cortically induced tingles to similar sensations produced in the more usual way, by applying a brief electrical pulse to the hand itself (Libet 1965; 1981; 1982; 1985a; Libet et al. 1979; see also Churchland 1981a; 1981b; Dennett 1979; Honderich 1984; Popper & Eccles 1977). He argued that although in each case it took considerable time (approximately 500 msec) to achieve "neuronal adequacy" (the stage at which cortical processes culminate to yield a conscious experience of a tingle), when the hand itself was stimulated, the experience was "automatically . . . referred backwards in time."

Most strikingly, Libet reported instances in which a subject's left *cortex* was stimulated *before* his left *hand* was stimulated, something one would tend to expect to give rise to two felt tingles: First right hand (cortically induced) and then left hand. In fact, however, the subjective report was reversed: "first left, then right." Even in cases of simultaneous stimulation, one might have thought, the left-hand tingle should be felt second, because of the additional distance (close to a meter) nerve impulses from the left hand must travel to the brain.

Libet interprets his results as raising a serious challenge to materialism: "A dissociation between the timings of the corresponding 'mental' and 'physical' events would seem to raise serious though not insurmountable difficulties for the . . . theory of psychoneural identity" (1979, p. 222). According to Eccles, this challenge cannot be met:

> This antedating procedure does not seem to be explicable by any neurophysiological process. Presumably it is a strategy that has been learnt by the self-conscious mind . . . the antedating sensory experience is attributable to the ability of the self-conscious mind to make slight temporal adjustments, i.e., to play tricks with time. (Popper & Eccles 1977, p. 364)

D. Subjective delay of consciousness of intention. In other experiments, Libet asked subjects to make "spontaneous" decisions to flex one hand at the wrist while noting the position of a revolving spot (the "second hand" on a clock, in effect) at the precise time they formed the intention (Libet 1985a; 1987; 1989a; see also the accompanying commentaries). Subjects' reports of these subjective simultaneities were then plotted against the timing of relevant electrophysiological events in their brains. Libet found evidence that these "conscious decisions" lagged between 350 and 400 msec behind the onset of "readiness potentials" he was able to record from scalp electrodes, which, he claims, tap the neural events that determine the voluntary actions performed. He concludes that "cerebral initiation of a spontaneous voluntary act begins unconsciously" (1985a, p. 529). That one's consciousness might lag behind the brain processes that control one's body seems to some an unsettling and even depressing prospect, ruling out a real (as opposed to illusory) "executive role" for "the conscious self." (See the discussions by many commentators in *BBS*: Eccles 1985; Mortenson 1985; Van Gulick 1985; and in Pagels 1988, pp. 233ff; and Calvin 1990, pp. 80–81. But see, for a view close to ours, Harnad 1982.)

In none of these cases would there be prima facie evidence of any anomaly were we to forego the opportunity to record the subjects' *verbal reports* of their experiences and subject them to semantic analysis. No sounds appear to issue from heads before lips move, nor do hands move before the brain events that purportedly cause them, nor do events occur in the cortex in advance of the stimuli that are held to be their source. Viewed strictly as the internal and external behavior of a biologically implemented control system for a body, the events observed and clocked in the experiments mentioned exhibit no apparent violations of everyday mechanical causation – of the sort to which Galilean/Newtonian physics provides the standard approximate model. Libet said it first: "It is important to realize that these subjective referrals and corrections are apparently taking place at the level of the *mental* 'sphere'; they are not apparent, as such, in the activities at neural levels" (1982, p. 241).

Put more neutrally (pending clarification of what Libet means by the "mental 'sphere'"), only through the subjects' verbalizations about their subjective experiences do we gain access to a perspective from which the anomalies can appear.[2] Once their verbalizations (including communicative button-pushes, etc.; Dennett 1982), are interpreted as a sequence of speech acts, their *content* yields a time series, *the subjective sequence of the stream of consciousness*. One can then attempt to put this series into registration with another time series, *the objective sequence of observed events in the environment and in the nervous system*. It is the apparent failures of registration, holding constant the assumption that causes precede their effects, that constitute the supposed anomalies (cf. Hoy 1982).

One could, then, "make the problems disappear" by simply refusing to take introspective reports seriously. Although some hearty behaviorists may cling comfortably to the abstemious principle, "Eschew content!" (Dennett 1978), the rest of us prefer to accept the challenge to make sense of what Libet calls "a primary phenomenological aspect of our human existence in relation to brain function" (1985a, p. 534).

> The reports by subjects about their different experiences . . . were not theoretical constructs but empirical observations. . . . The method of introspection may have its limitations, but it can be used appropriately within the framework of natural science, and it is absolutely essential if one is trying to get some experimental data on the mind-brain problem. (Libet 1987, p. 785)

In each example an apparent dislocation in time threatens the *prima facie* plausible thesis that our conscious perceptions are caused by events in our nervous systems, and our conscious acts, in turn, cause events in our nervous systems that control our bodily acts. To first appearances, the anomalous phenomena show that these two standard causal links cannot be sustained unless we abandon a foundational – some would say a logically necessary – principle: *Causes precede their effects*. It seems that in one case (subjective delay of awareness of intention), our conscious intentions *occur too late* to be the causes of their bodily expressions or implementations, and in the other cases, percepts *occur too early* to have been caused by their stimuli. The vertiginous alternative, that something in the brain (or "conscious self") can "play tricks with time" by "projecting" mental events backwards in time, would require us to abandon the foundational principle that causes precede their effects.

There is a widespread conviction that no such revolutionary consequence follows from any of these phenomena, a conviction we share. But some of the influential arguments that have been offered in support of this conviction persist in a commitment to the erroneous presuppositions that made the phenomena appear anomalous in the first place. These presuppositions are all the more insidious because although in their overt, blatant forms they are roundly disowned by one and all, they creep unnoticed back into place, distorting analysis and blinding theory-builders to other explanations.

2. The models in action: Diagnosing the tempting errors

2.1. The representation of temporal properties versus the temporal properties of representations. The brain, as the control system responsible for solving a body's real-time problems of interaction with the environment, is under significant time pressure. It must often arrange to modulate its output in light of its input within a time window that leaves no slack for delays. In fact, many acts can be only *ballistically* initiated; there is no time for feedback to adjust the control signals. Other tasks, such as speech perception, would be beyond the physical limits of the brain's machinery if they did not use ingenious anticipatory strategies that feed on redundancies in the input (Libermann 1970).

How, then, does the brain keep track of the temporal information it manifestly needs? Consider the following problem: Because the toe-brain distance is much greater than the hip-brain distance, or the shoulder-brain distance or the forehead-brain distance, stimuli delivered simultaneously at these different sites will arrive at Headquarters in staggered succession, if travel-speed is constant along all paths. How (one might be tempted to ask) does the brain "ensure central simultaneity of representation for distally simultaneous stimuli"? This encourages one to hypothesize some "delay loop" mechanism that could store the early arrivers until they could be put "in synch" with the latecomers, but this is a mistake. The brain should not solve *this* problem, for an obvious engineering reason: It squanders precious time by committing the full range of operations to a "worst case" schedule. Why should important signals from the forehead (for instance) dawdle in the anteroom just because there might someday be an occasion when concurrent signals from the toes need to be compared to (or "bound to") them?

The brain sometimes uses "buffer memories" to cushion the interface between its internal processes and the asynchronous outside world (Neisser 1967; Newell et al. 1989; Sperling 1960), but there are also ways for the brain to use the temporal information it needs without the delays required for imposing a master synchrony. The basic design principle is well illustrated in an example in which a comparable problem is confronted and (largely) solved, though on a vastly different temporal and spatial scale.

Consider the communication difficulties faced by the far-flung British Empire before the advent of radio and telegraph, as illustrated by the Battle of New Orleans. On January 8, 1815, 15 days after the truce was signed in Belgium, more than a thousand British soldiers were killed in this needless battle. We can use this debacle to see how the system worked. Suppose on Day 1 the treaty is signed in Belgium, with the news sent by land and sea to America, India, Africa. On Day 15 the battle is fought in New Orleans, and news of the defeat is sent by land and sea to England, India, and so on. On Day 20, too late, the news of the treaty (and the order to surrender) arrives in New Orleans. On Day 35, let's suppose, the news of the defeat arrives in Calcutta, but the news of the treaty doesn't arrive there until Day 40 (via a slow overland route). To the commander-in-chief in Calcutta, the battle would "seem" to have been fought before the treaty was signed – were it not for the practice of dating letters, which permits him to make the necessary correction.

These communicators solved their problems of communicating information about time by embedding representations of the relevant time information in the *content* of their signals, so that the arrival time of the signals themselves was *strictly irrelevant* to the information they carried. A date written at the head of a letter (or a dated postmark on the envelope) gives the recipient information about when it was sent, information that survives any delay in arrival.[3] This distinction between time represented (by the postmark) and time of representing (the day the letter arrives) is an instance of a familiar distinction between content and vehicle, and although the details of this particular solution are not available to the brain's communicators (because they don't "know the date" when they send their messages), the general principle of the content/vehicle distinction is relevant to information-processing models of the brain in ways that have not been well appreciated.[4]

In general, we must distinguish features of representings from the features of representeds (Neumann 1990); someone can shout "softly, on tiptoe" at the top of his lungs, there are gigantic pictures of microscopic objects and oil paintings of artists making charcoal sketches. The top sentence of a written description of a standing man need not describe his head, nor the bottom sentence his feet. To suppose otherwise is confusedly to superimpose two different spaces: The representing space and the represented space. The same applies to time. Consider the *spoken phrase*, "a bright, brief flash of red light." The beginning of *it* is "a bright" and the end of it is "red light." Those portions of that speech event are not themselves representations of onsets or terminations of a brief red flash (cf. Efron 1967, p. 714). No informing event in the nervous system can have zero duration (any more than it can have zero spatial extent), so it has an onset and termination separated by some amount of time. If it *represents* an event in experience, then the event it represents must itself have nonzero duration, an onset, a middle, and a termination. But there is no reason to suppose that the beginning of the representing represents the beginning of the represented.[5]

Similarly, the representing by the brain of "A before B" does not have to be accomplished by first:

a representing of A,

followed by:

a representing of B.

"B after A" is an example of a (spoken) vehicle that

represents A as being before B, and the brain can avail itself of the same freedom of temporal placement. What matters for the brain is not necessarily *when* individual presenting events happen in various parts of the brain (as long as they happen in time to control the things that need controlling!) but their *temporal content*. That is, what matters is that the brain can proceed to control events "under the assumption that A happened before B" whether or not the information that A has happened enters the relevant system of the brain and gets recognized as such before or after the information that B has happened. (Recall the commander-in-chief in Calcutta: First he is informed of the battle, and then he is informed of the truce, but because he can extract from this the information that the truce came first, he can act accordingly.) Systems in various locations in the brain can, in principle, avail themselves of similar information-processing, and that is why fixing the exact time of onset of some representing element in some place in the brain does not provide a temporal landmark relative to which other elements *in the subjective sequence* can – or must – be placed.

How are temporal properties really inferred by the brain? Systems of "date stamps" or "postmarks" are not theoretically impossible (Glynn 1990), but there is a cheaper, less foolproof but biologically more plausible way: by what we might call *content-sensitive settling*. A useful analogy would be the film studio where the sound track is "synchronized" with the film. The various segments of audio tape may by themselves have lost all their temporal markers, so that there is no simple, mechanical way of putting them into apt registration with the images. But sliding them back and forth relative to the film and looking for convergences, will usually swiftly home in on a "best fit." The slap of the slateboard at the beginning of each take provides a double saliency, an auditory and a visual clap, to slide into synchrony, pulling the rest of the tape and the frames into position at the same time. But there are typically so many points of mutually salient correspondence that this conventional saliency at the beginning of each take is just a handy redundancy. Getting the registration right depends on the *content* of the film and the tape, but not on sophisticated analysis of the content. An editor who knew no Japanese would find synchronizing a Japanese soundtrack to a Japanese film difficult and tedious but not impossible. Moreover, the temporal order of the stages of the process of putting the pieces into registration is independent of the content of the product; the editor can organize scene three before organizing scene two, and in principle could even do the entire job running the segments "in reverse."

Quite "stupid" processes can do similar jiggling and settling in the brain. The computation of depth in random-dot stereograms (Julesz 1971) is a spatial problem for which we can readily envisage temporal analogues. If the system receives stereo pairs of images, the globally optimal registration can be found without first having to subject each data array to an elaborate process of feature extraction. There are enough lowest-level coincidences of saliency – the individual dots in a random dot stereogram – to dictate a solution. In principle, then, the brain can solve some of its problems of temporal inference by such a process, drawing data not from left and right eyes, but from whatever information-sources are involved in a process requiring temporal judgments. (See Gallistel, 1990, especially pp. 539–49, for a discussion of the requirements for "spatiotemporal specification.")

Two important points follow from this. First, such temporal inferences can be drawn (such temporal discriminations can be made) by comparing the (low-level) *content* of several data arrays, and this real time process need not occur in the temporal order that its product eventually represents. Second, once such a temporal inference has been drawn, which may be *before* high-level features have been extracted by other processes, it does not have to be drawn again! There does not have to be a *later* representation in which the high-level features are "presented" in a real time sequence for the benefit of a second sequence-judger. In other words, having drawn inferences from these juxtapositions of temporal information, the brain can go on to represent the results in any format that fits its needs and resources – not necessarily a format in which "time is used to represent time."

There remains a nagging suspicion that whereas the brain may take advantage of this representational freedom for other properties, it cannot do so for the property of temporal sequence. Mellor explicitly enunciates this assumption, deeming it too obvious to need support:

> Suppose for example I see one event *e* precede another, *e**. I must first see *e* and then *e**, my seeing of *e* being somehow recollected in my seeing of *e**. That is, my seeing of *e* affects my seeing of *e**: This is what makes me – rightly or wrongly – see *e* precede *e**rather than the other way round. But seeing *e* precede *e** means seeing *e* first. So the causal order of my perceptions of these events, by fixing the temporal order I perceive them to have, fixes the temporal order of the perceptions themselves. . . . the striking fact . . . should be noticed, namely that perceptions of temporal order need temporally ordered perceptions. *No other property or relation has to be thus embodied in perceptions of it* [our emphasis]: perceptions of shape and colour, for example, need not themselves be correspondingly shaped or coloured. (Mellor 1981, p. 8)

We believe this is false, but there is something right about it. Because the fundamental function of representation in the brain is to control behavior in real time, the timing of representings is *to some degree* essential to their task, in two ways. First, the timing may, at the outset of a perceptual process, be *what determines the content*. Consider how to distinguish a spot moving from right to left from a spot moving from left to right on a motion picture screen. The *only* difference between the two may be the temporal order in which two frames (or more) are projected. If the brain determines "first A, then B" the spot is seen as moving in one direction; if the brain determines "first B, then A" the spot is seen as moving in the opposite direction. This discrimination is, then, as a matter of logic, based on the brain's capacity to make a temporal order judgment of a particular level of resolution. Motion picture frames are usually exposed at the rate of 24 per second, and so the visual system can resolve order between stimuli that occur within about 50 msec. This means that the actual temporal properties of signals – their onset times, their velocity in the system, and hence their arrival times – must be accurately controlled until such a discrimination is made. But once it is made locally by some circuit in the visual system (even as peripherally

as the ganglion cells of the rabbit's retina! – Barlow & Levick 1965), the content "from left to right" can then be sent, in a temporally sloppy way, anywhere in the brain where this directional information might be put to use. This way one can explain the otherwise puzzling fact that at interstimulus intervals at which people are unable to perform above chance on temporal order judgments, they perform flawlessly on other judgments that logically call for the same temporal acuity. Thus Efron (1973) showed that subjects could easily distinguish sounds, flashes, and vibrations that differed only in the order in which two component stimuli occurred at a fraction of the interstimulus interval at which they can explicitly specify their order.

A second constraint on timing has already been noted parenthetically above: It does not matter in what order representations occur so long as they occur in time to contribute to the control of the appropriate behavior. The function of a representing may depend on meeting a *deadline*, which is a temporal property of the vehicle doing the represent*ing*. This is particularly evident in such time-pressured environments as the imagined Strategic Defense Initiative. The problem is not how to make computer systems represent, accurately, missile launches, but how to represent a missile launch accurately during the brief time while one can still do something about it. A message that a missile was launched at 6:04:23.678 A.M. EST may accurately represent the time of launch forever, but its utility may utterly lapse at 6:05 A.M. EST. For any task of control, then, there is a *temporal control window* within which the temporal parameters of representings may in principle be moved around ad lib.

The deadlines that limit such windows are not fixed, but rather depend on the task. If, rather than intercepting missiles, you are writing your memoirs or answering questions at the Watergate hearings (Neisser 1981), you can recover the information you need about the sequence of events in your life to control your actions in almost any order, and you can take your time drawing inferences.

These two factors explain what is plausible in Mellor's claim, without supporting the invited conclusion that all perceptions of temporal order must be accomplished in a single place by a process that observes *seriatim* a succession of "perceptions" or other representations. Once the perceptual processes *within* an observer have begun to do their work, providing the necessary discriminations, there is no point in undoing their work to provide a job for a yet more interior observer.

Causes must precede effects. This fundamental principle ensures that temporal control windows are bounded at both ends: by the earliest time at which information could arrive in the system, and by the latest time at which information could contribute causally to the control of a particular behavior. Moreover, the principle applies to the multiple distributed processes that achieve such control. Any particular process that requires information from some source must indeed wait for that information; it can't get there till it gets there. This is what rules out "magical" or precognitive explanations of the color-switching phi phenomenon, for example. The content *green spot* cannot be attributed to any event, conscious or unconscious, until the light from the green spot has reached the eye and triggered the normal neural activity in the visual system up to the level at which the discrimination of green is accomplished. Moreover, all content reported or otherwise expressed in subsequent behavior must have been "present" (in the relevant place in the brain, but not necessarily in consciousness) in time to have contributed causally to that behavior. For instance, if a subject in an experiment *says* "dog" in response to a visual stimulus, we can work backwards from the behavior, which was clearly controlled by a process that had the content *dog* (unless the subject says "dog" to every stimulus, or spends the day saying "dog dog dog . . . " etc.) And since it takes on the order of 100 msec to execute a speech intention of this sort, we can be quite sure that the content *dog* was present in (roughly) the language areas of the brain by 100 msec before the utterance. Working from the other end, we can determine the earliest time the content *dog* could have been computed or extracted by the visual system from the retinal input, and even, perhaps, follow its creation and subsequent trajectory through the visual system and into the language areas.

What would be truly anomalous (indeed a cause for lamentations and the gnashing of teeth) would be if the time that elapsed between the *dog*-stimulus and the "dog"-utterance were less than the time physically required for this content to be established and moved through the system. No such anomalies have been uncovered, however. It is only when we try to put the sequence of events thus detectable in the objective processing stream into registration with the subject's subjective sequence *as indicated by what the subject subsequently says* that we have any sign of anomaly at all.

2.2. Orwellian and Stalinesque revisions: The illusion of a distinction. Now let us see how the two different models, the Cartesian Theater and Multiple Drafts, deal with the presumed anomalies, starting with the simpler and less controversial phenomena. The Cartesian Theater model postulates a place within the brain where what happens "counts"; that is, it postulates that the features of events occurring within this functionally definable boundary (whatever it is) are definitive or constitutive features of conscious experience. (The model applies to all features of subjective experience, but we are concentrating on temporal features.) This implies that all revisions of content accomplished by the brain can be located relative to this place, a deeply intuitive – but false – implication that can be illustrated with a thought experiment.

Suppose we tamper with your brain, inserting in your memory a bogus woman wearing a hat where none was (e.g., at the party on Sunday). If on Monday, when you recall the party, you remember her, and can find no internal resources for so much as doubting the veracity of your memory, we could all agree that you never *did* experience her; that is, not at the party on Sunday. Of course your subsequent experience of (bogus) recollection can be as vivid as may be, and on Tuesday we can certainly agree that you have had vivid conscious experiences of there being a woman in a hat at the party, but the *first* such experience, we would insist, was on Monday, not Sunday (although it doesn't seem this way to you).

We lack the power to insert bogus memories by neurosurgery, but sometimes our memories play tricks on us, so what we cannot yet achieve surgically happens in the

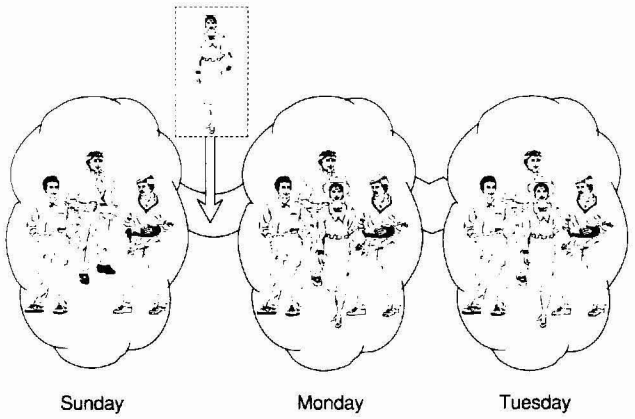

Figure 1. Post-experiential memory tampering.

brain on its own. Sometimes we seem to remember, even vividly, experiences that never occurred. We might call such post-experiential contaminations or revisions of memory *Orwellian*, recalling George Orwell's chilling vision of the Ministry of Truth in *1984*, which busily rewrote history and thus denied access to the (real) past to all who followed.

Orwellian revision is one way to fool posterity. Another is to stage show trials, carefully scripted presentations of false testimony and bogus confessions, complete with simulated evidence. We might call this ploy *Stalinesque*. Notice that if we are usually sure which mode of falsification has been attempted on us, the Orwellian or the Stalinesque, this is just a happy accident. In any *successful* disinformation campaign, were we to wonder whether the accounts in the newspapers were Orwellian accounts of trials that never happened at all, or true accounts of phony show trials that actually did happen, we might be unable to tell the difference. If *all* the traces – newspapers, videotapes, personal memoirs, inscriptions on gravestones, living witnesses, and so on – have been either obliterated or revised, we will have no way of knowing which sort of fabrication happened: a fabrication *first*, culminating in a staged trial whose accurate history we now have before us, or *after* a summary execution, history-fabrication covering up the deed. No trial of any sort *actually* took place.

The distinction between reality and (subsequent) appearance, and the distinction between Orwellian and Stalinesque methods of producing misleading archives, work unproblematically in the everyday world, at macroscopic time scales. One might well think these distinctions apply unproblematically *all the way in*. That is the habit of thought that produces the cognitive illusion of Cartesian materialism. We can catch it in the act in a thought experiment that differs from the first one in nothing but time scale.

Suppose a long-haired woman jogs by. About one second *after* this, a subterranean memory of some earlier woman – a short-haired woman with glasses – contaminates the memory of what you have just seen: When asked a minute later for details of the woman you just saw, you report, sincerely but erroneously, that she was wearing glasses. Just as in the previous case, we are inclined to say that your original *visual* experience, as opposed to the memory of it seconds later, was *not* of a woman with

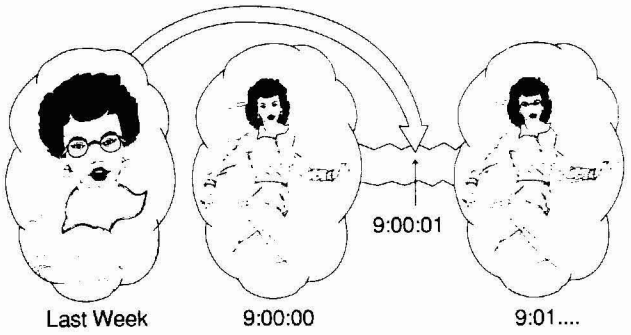

Figure 2. Orwellian revision.

glasses. But because of the subsequent memory-contaminations, it seems to you exactly as if at the first moment you saw her, you were struck by her eyeglasses. An Orwellian, postexperiential revision has happened: There was a fleeting instant, before the memory contamination took place, when it *didn't* seem to you she had glasses. For that brief moment, the *reality* of your conscious experience was a long-haired woman *without* eyeglasses, but this historical fact has become inert; it has left no trace, thanks to the contamination of memory that came one second after you glimpsed her.

This understanding of what happened is jeopardized by an alternative account, however. Your subterranean earlier memories of that short-haired woman with the glasses could just as easily have contaminated your experience *on the upward path*, in the processing of information that occurs "prior to consciousness" so that you actually *hallucinated* the eyeglasses from the very beginning of your experience. In that case, your obsessive memory of the woman with glasses would be playing a Stalinesque trick on you, creating a "show trial" for you to experience, which you then accurately recall at later times, thanks to the record in your memory. To naive intuition these two cases are as different as can be. Told the first way (Figure 2), you suffer no hallucination at the time the woman jogs by, but suffer subsequent memory-hallucinations: You have false memories of your actual ("real") experience. Told the second way (Figure 3), you hallucinate when she runs by, and then accurately remember that hallucination (which "really did happen in consciousness") thereafter. Surely these are distinct possibilities, no matter how finely we divide up time?

No. Here the distinction between perceptual revisions and memory revisions that works so crisply at other scales is not guaranteed application. We have moved into the

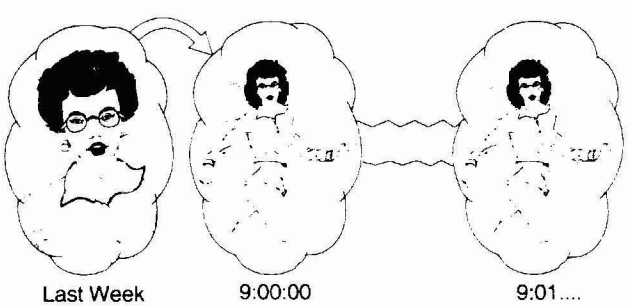

Figure 3. Stalinesque show trial.

foggy area in which the subject's point of view is spatially and temporally smeared, and the question *Orwellian or Stalinesque?* (post-experiential or pre-experiential) need have no answer. The boundary between perception and memory, like most boundaries between categories, is not perfectly sharp, as has often been noted.

There is a time window that began when the long-haired woman jogged by, exciting your retinas, and ended when you expressed – to yourself or someone else – your eventual conviction that she was wearing glasses. At some time during this interval, the content *wearing glasses* was spuriously added to the content *long-haired woman*. We may assume (and might eventually confirm in detail) that there was a brief time when the content *long-haired woman* had already been discriminated in the brain but *before* the content *wearing glasses* had been erroneously "bound" to it. Indeed, it would be plausible to suppose that this discrimination of a long-haired woman was what triggered the memory of the earlier woman with the glasses. What we would not know, however, is whether this spurious binding was before or after the fact – the presumed fact of "actual conscious experience." Were you first conscious of a long-haired woman without glasses and then conscious of a long-haired woman with glasses, a subsequent consciousness that wiped out the memory of the earlier experience, or was the very first instant of conscious experience already spuriously tinged with eyeglasses? If Cartesian materialism were correct, this question would have to have an answer, even if we – and you – could not determine it retrospectively by any test, for the content that "crossed the finish first" was either *long-haired woman* or *long-haired woman with glasses*. But what happens to this question if Cartesian materialism is incorrect (as just about everyone agrees)? Can the distinction between pre-experiential and post-experiential content revisions be maintained?

An examination of the color phi phenomenon shows that it cannot. On the first trial (i.e., without conditioning), subjects *report* seeing the color of the moving spot switch in midtrajectory from red to green – a report sharpened by Kolers's ingenious use of a pointer device which subjects retrospectively-but-as-soon-as-possible "superimposed" on the trajectory of the illusory moving spot; such pointer locations had the content: "The spot changed color right about *here*" (Kolers & von Grünau 1976, p. 330). Recall Goodman's (1978, p. 73) expression of the puzzle: "How are we able . . . to fill in the spot at the intervening place-times along a path running from the first to the second flash *before that second flash occurs*?"

Consider, first, a Stalinesque mechanism: In the brain's editing room, located before consciousness, there is a delay, a loop of slack like the "tape delay" used in broadcasts of "live" programs, which gives the censors in the control room a few seconds to bleep out obscenities before broadcasting the signal. *In the editing room*, first frame A, of the red spot, arrives, and then, when frame B, of the green spot, arrives, some interstitial frames (C and D) can be created and then spliced into the film (in the order A, C, D, B) on its way to projection in the theater of consciousness. By the time the "finished product" arrives at consciousness, it already has its illusory insertion.

Alternatively, there is the hypothesis of an Orwellian mechanism: Shortly after the awareness of the first spot *and* the second spot (with no illusion of apparent motion at all), a revisionist historian of sorts, in the brain's memory-library receiving station, notices that the unvarnished history of this incident doesn't make enough sense, so he "interprets" the brute events, red-followed-by-green, by making up a narrative about the intervening passage, complete with midcourse color change, and installs this history, incorporating his glosses, frames C and D (in Figure 4), in the memory library for all future reference. Because he works fast, within a fraction of a second – the amount of time it takes to frame (but not utter) a verbal report of what you have experienced – the record you rely on, stored in the library of memory, is already contaminated. You *say* and *believe* that you saw the illusory motion and color change, but that is really a memory hallucination, not an accurate recollection of your original awareness.

How could we see which of these hypotheses is correct? It might seem that we could rule out the Stalinesque hypothesis quite simply, because of the delay in consciousness it postulates. In Kolers and von Grünau's experiment, there was a 200 msec difference in onset between the red and green spot, and since, *ex hypothesi*, the *whole experience* cannot be composed by the editing room until after the content *green spot* has reached the editing room, consciousness of the initial red spot will have to be delayed by at least that much. (If the editing room sent the content *red spot* up to the theater of consciousness immediately, before receiving frame B and then fabricating frames C and D, the subject would presumably experience a gap in the film, a noticeable delay of around 200 msec between A and C.)

Suppose we ask subjects to press a button "as soon as you experience a red spot." We would find little or no difference in response time to a red spot alone versus a red spot followed 200 msec later by a green spot (in which case the subjects report color-switching apparent motion). This could be because there is *always* a delay of at least 200 msec in consciousness, but aside from the biological implausibility of such a squandering of time, there is the evidence from many quarters that responses under conscious control, although slower than such responses as reflex blinks, occur with close to the minimum latencies that are physically possible; after subtracting the demonstrable travel times for incoming and outgoing pulse trains, and the response preparation time, there is little time left over in "central processing" in which to hide a 200 msec delay. So the responses had to have been

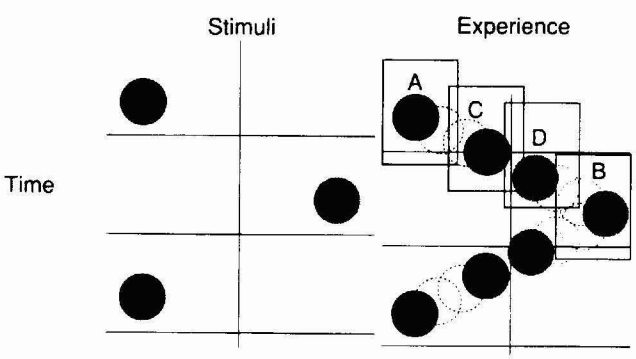

Figure 4. Frames C and D inserted in the editing room.

initiated before the discrimination of the second stimulus, the green spot. This would seem overwhelmingly to favor the Orwellian, post-experiential mechanism: As soon as the subject *becomes conscious* of the red spot, he initiates a button-press. *While that button press is forming*, he becomes conscious of the green spot. *Then* both these experiences are wiped from memory, replaced in memory by the revisionist record of the red spot moving over and then turning green halfway across. He readily and sincerely (but mistakenly) reports having seen the red spot moving toward the green spot before changing color.

If the subject were to insist that he really was conscious from the very beginning of the red spot moving and changing color, the Orwellian theorist would firmly explain to him that he is wrong; his memory is playing tricks on him; the fact that he pressed the button when he did is conclusive evidence that he was conscious of the (stationary) red spot before the green spot had even occurred. After all, his instructions were to press the button *when he was conscious of* a red spot. He must have been conscious of the red spot about 200 msec before he could have been conscious of it moving and turning green. If that is not how it seems to him, he is simply mistaken.

The defender of the Stalinesque (pre-experiential) alternative is not defeated by this, however. Actually, he insists, the subject responded to the red spot *before* he was conscious of it! The directions to the subject (to respond to a red spot) had somehow trickled down from consciousness into the editing room, which *unconsciously* initiated the button-push before sending the edited version (frames ACDB) up to consciousness for "viewing." The subject's memory has played no tricks on him; he is reporting exactly what he was conscious of, unless he insists that he pushed the button after consciously seeing the red spot; his "premature" button-push was unconsciously (or preconsciously) triggered (cf. Velmans 1991).

Where the Stalinesque theory postulates a button-pushing reaction to an *un*conscious detection of a red spot, the Orwellian theory postulates a *conscious* experience of a red spot that is immediately obliterated from memory by its sequel. So here is the rub: We have two different models of what happens in the phi phenomenon: one posits a Stalinesque "filling in" on the upward, pre-experiential path, and the other posits an Orwellian "memory revision" on the downward, post-experiential path, and *both* of them are consistent with *whatever* the subject says or thinks or remembers. Note that the inability to distinguish these two possibilities does not apply only to the *outside observers* who might be supposed to lack some private data to which the subject had "privileged access." You, as a subject in a phi phenomenon experiment, *could not* discover anything in the experience from your own first-person perspective that would favor one theory over the other; the experience would "feel the same" on either account. As the interstimulus interval is lengthened subjects pass from seeing apparent motion to seeing individual stationary flashes. There is an intermediate range of intervals where the phenomenology is somewhat paradoxical: You see the spots as two stationary flashes *and* as one thing moving. This sort of apparent motion is readily distinguishable from the swifter, smoother sort of apparent motion of cinema, for instance, but your capacity to make *this* discrimination is not relevant to the dispute between the Orwellian and the Stalinesque theorist. They agree that you can make this discrimination under the right conditions; what they disagree about is how to describe the cases of apparent motion that you *can't* tell from real motion – the cases in which you really (mis-)*perceive* the illusory motion. To put it loosely, in these cases is your memory playing tricks with you, or are just your eyes playing tricks with you? You can't tell "from the inside."

We can see the same indistinguishability even more clearly when we see how the two different models handle the well-studied phenomenon of *metacontrast* (for a review, see Breitmeyer 1984). If a stimulus is flashed briefly on a screen and then followed, after a brief interstimulus interval, by a second "masking" stimulus, subjects *report* seeing only the second stimulus. (And if you put yourself in the subject's place you will see for yourself; you will be prepared to swear that there was only one flash.) The standard description of such phenomena is that the second stimulus somehow *prevents conscious experience* of the first stimulus (in other words, it somehow waylays the first stimulus on its way to consciousness). But people can nevertheless do much better than chance if required to guess whether there were two stimuli. This only shows once again that stimuli can have their effects on us without our being conscious of them. This standard line is, in effect, the Stalinesque model of metacontrast: The first stimulus never gets to play on the stage of consciousness; it has whatever effects it has entirely unconsciously. But we have just uncovered a second, Orwellian model of metacontrast: Subjects are indeed conscious of the first stimulus (which would "explain" their capacity to guess correctly) but their memory of this conscious experience is almost entirely obliterated by the second stimulus (which is why they deny having seen it, in spite of their tell-tale better-than-chance guesses).[6]

Both the Orwellian and the Stalinesque version of the Cartesian Theater model can deftly account for *all* the data – not just the data we already have, but the data we can imagine getting in the future. They both account for the verbal reports: One theory says they are innocently mistaken whereas the other says they are accurate reports of experienced "mistakes." (A similar verdict is suggested in the commentaries of Holender 1986; see especially Dixon 1986; Erdelyi 1986; Marcel 1986; Merikle & Cheesman 1986.) They agree about just where in the brain the mistaken content enters the causal pathways; they just disagree about whether that location is pre-experiential or post-experiential. They both account for the nonverbal effects: One says they are the result of unconsciously discriminated contents while the other says they are the result of consciously discriminated but forgotten contents. They agree about just where and how in the brain these discriminations occur; they just disagree about whether to interpret those processes as happening inside or outside the charmed circle of consciousness. Finally, they both account for the subjective data – whatever is obtainable "from the first-person-perspective" – because they agree about how it ought to "feel" to subjects: Subjects should be unable to tell the difference between misbegotten experiences and immediately misremembered experiences. So, in spite of first appearances, there is really only a verbal difference between the two theories (cf. Reingold & Merikle 1990). They tell exactly the same story except for where they

place a mythical Great Divide, a point in time (and hence a place in space) whose *fine-grained* location is nothing that subjects can help them locate, and whose location is also neutral with regard to all other features of their theories. This is a difference that makes no difference.

Consider a contemporary analogy. With the advent of word-processing and desktop publishing and electronic mail, we are losing the previously quite hard-edged distinction between pre-publication editing, and post-publication correction of "errata." With multiple drafts in electronic circulation, and with the author readily making revisions in response to comments received by electronic mail, calling one of the drafts the canonical text – the text of "record," the one to cite in one's own publications – becomes a somewhat arbitrary matter. Often most of the intended readers, the readers whose reading of the text matters, read only an early draft; the "published" version is archival and inert. If it is important effects we are looking for, then, most if not all the important effects of writing a text are now spread out over many drafts, not postponed until after publication. It used to be otherwise; virtually all of a text's important effects happened *after* appearance in a book or journal and *because of* its making such an appearance. All the facts are in, and now that the various candidates for the "gate" of publication can be seen no longer to be functionally important, if we feel we need the distinction at all, we will have to decide arbitrarily what is to count as publishing a text. There is no natural summit or turning point in the path from draft to archive.

Similarly – and this is the fundamental implication of the Multiple Drafts model – if one wants to settle on some moment of processing in the brain as the moment of consciousness, this has to be arbitrary. One can always "draw a line" in the stream of processing in the brain, but there are no functional differences that could motivate declaring all prior stages and revisions unconscious or preconscious adjustments and all subsequent emendations to the content (as revealed by recollection) to be post-experiential memory-contamination. The distinction lapses at close quarters.

Another implication of the Multiple Drafts model, in contrast to the Cartesian Theater, is that there is no need – or room – for the sort of "filling in" suggested by frames C and D of Figure 4. Discussing Kolers' experiment, Goodman notes that it "seems to leave us a choice between a retrospective construction theory and a belief in clairvoyance" (1978, p. 83). What then is "retrospective construction"?

> Whether perception of the first flash is thought to be *delayed or preserved or remembered* [our emphasis], I call this the retrospective construction theory – the theory that the construction perceived as occurring between the two flashes is accomplished not earlier than the second.

It seems at first that Goodman does not choose between a Stalinesque theory (perception of the first flash is delayed) and an Orwellian theory (the perception of the first flash is preserved or remembered), but his Orwellian revisionist does not merely adjust judgments; he *constructs* material to *fill in* the gaps: "Each of the intervening places along a path between the two flashes is filled in . . . with one of the flashed colors rather than with successive intermediate colors" (Goodman 1978, p. 85). What Goodman overlooks is the possibility that the brain doesn't actually have to go to the trouble of "filling in" anything with "construction," for no one is looking. As the Multiple Drafts model makes explicit, once a discrimination has been made once, it does not have to be made again; the brain just adjusts to the conclusion that is drawn, making the new interpretation of the information available for the modulation of subsequent behavior. Recall the commander-in-chief in Calcutta; he just had to *judge* that the truce came before the battle; he didn't also have to mount some sort of pageant of "historical reconstruction" to watch, in which he receives the letters in the "proper" order.

Similarly, when Goodman (1978) proposes that "the intervening motion is produced retrospectively, built only after the second flash occurs, and projected backwards in time," this suggests ominously that a final film is made and then run through a magical projector whose beam somehow travels backwards in time onto the mind's screen. Whether or not this is just what Van der Waals and Roelofs (1930) had in mind when they proposed "retrospective construction," it is presumably what led Kolers (1972, p. 184) to reject their hypothesis, insisting that all construction is carried out in "real time." Why, though, should the brain bother to "produce" the "intervening motion"? Why not just conclude that there was intervening motion, and encode that "retrospective" content into the processing stream? This would suffice for it to seem to the subject that intervening motion had been experienced.

Our Multiple Drafts model agrees with Goodman that retrospectively the brain creates the content (the judgment) that there was intervening motion, and this content is then available to govern activity and leave its mark on memory. But our model claims that the brain does not bother "constructing" any representations that go to the trouble of "filling in" the blanks. That would be a waste of time and (shall we say?) *paint*. The judgment is *already in*, so the brain can get on with other tasks![7]

Goodman's "projection backwards in time," like Libet's "backwards referral in time," is an equivocal phrase. It might mean something modest and defensible: A *reference to some past time* is included in the content. On this reading it could be a claim like, "This novel takes us back to ancient Rome," which almost no one would interpret in a metaphysically extravagant way, as claiming that the novel was some sort of time travel machine. This is the reading that is consistent with Goodman's other views, but Kolers apparently took it to mean something metaphysically radical: that there was some actual projection of one thing at one time to another time. As we shall see, the same equivocation bedevils Libet's interpretation of his phenomena.

The model of the Cartesian Theater creates artifactual puzzle questions that cannot be answered, whereas for our model these questions cannot meaningfully arise. This can be seen by applying both models to other experiments that probe the limits of the distinction between perception and memory. A normally sufficient, but not necessary, condition for having experienced something is subsequent verbal report, and this is the anchoring case around which all the puzzle cases revolve. Suppose that although one's brain has registered – that is, responded to – (some aspects of) an event, something intervenes between that internal response and a subse-

quent occasion for verbal report. If there was no time or opportunity for an initial overt response of any sort, and if the intervening events prevent later overt responses (verbal or otherwise) from incorporating reference to some aspect(s) of the first event, this creates a puzzle question: Were they never consciously perceived, or have they been rapidly forgotten?

Consider the familiar span of apprehension. Multiple letters are simultaneously briefly exposed. Some are identified. The rest were certainly seen. The subject insists they were there, knows their number, and has the impression that they were clearcut and distinct. Yet he cannot identify them. Has he failed "really" to perceive them, or has he rapidly "forgotten" them? Or consider an acoustic memory span test, administered at a rapid rate, for example, 4 items a second, so that the subject perforce cannot respond till the acoustic event is over. He identifies some, not others. Yet, subjectively he heard all of them clearly and equally well. Did he not genuinely perceive or did he forget the rest?

And if, under still more constricted circumstances such as metacontrast, the subject even lacks all conviction that the unrecallable items *were there*, should we take this judgment as conclusive grounds for saying he did not experience them, even if they prove to have left other contentful traces on his subsequent behavior? If there is a Cartesian Theater, these questions demand answers, because what gets into the theater and when is supposedly determinate, even if the boundaries appear fuzzy because of human limitations of perception and memory.

Our Multiple Drafts model suggests a different perspective on these phenomena. When a lot happens in a short time, the brain may make simplifying assumptions (for a supporting view, see Marcel 1983). In metacontrast, the first stimulus may be a disc and the second stimulus a ring that fits closely outside the space where the disc was displayed. The outer contour of a disc rapidly turns into the inner contour of a ring. The brain, initially informed just that something happened (something with a circular contour in a particular place), swiftly receives confirmation that there was indeed a ring, with an inner and outer contour. Without further supporting evidence that there was a disc, the brain arrives at the conservative conclusion that there was only a ring. Should we insist that the disc was experienced because *if the ring hadn't intervened* the disc would have been reported? Our model of how the phenomenon is caused shows that there is no motivated way of settling such border disputes: Information about the disc was briefly in a functional position to contribute to a later report, but this state lapsed; there is no reason to insist that this state was inside the charmed circle of consciousness until it got overwritten, or contrarily, to insist that it never quite achieved this state. Nothing discernible to "inside" or "outside" observers could distinguish these possibilities.

In color phi, the processes that calculate that the second spot is green and that there is motion proceed roughly simultaneously (in different parts of the brain) and eventually contribute to the process that concludes that the red spot moved over and abruptly turned green on the way. That conclusion is achieved swiftly enough, in the standard case, to overwhelm or replace any competing contents before they can contribute to the framing of a report. So the subject says – and believes – just what Kolers and von Grünau report, *and that is what the subject was conscious of.* Was the subject *also* conscious a fraction of a second earlier of the stationary red spot? Ask him. If the interstimulus interval is made somewhat longer, there will come a point where the subject *does* report an experience of first a stationary red spot, then a green spot, and then a *noticeably retrospective* sense that the red spot ("must have") moved over and changed color. This experience has – as the subject will tell you – a quite different phenomenology. Apparent motion is experienced under such conditions, but it is obviously different from ordinary motion, and from swifter varieties of apparent motion. How is it different? The subject notices the difference! In this case it does seem to him as if he only later "realized" that there had been motion. But in cases in which this retrospective element is lacking it is still the case that the discrimination of motion-with-color-change is achieved after the colors and locations of the spots were discriminated – and there is no later process of "filling in" required.

In the cutaneous "rabbit," the shift in space (along the arm) is recorded over time by the brain. The number of taps is also recorded. Although in physical reality the taps were clustered at particular locations, the simplifying assumption is that they were distributed regularly across the space-time extent of the experience. The brain relaxes into this parsimonious though mistaken interpretation *after* the taps are registered, and this has the effect of wiping out earlier (partial) interpretations of the taps, but some side effects of those interpretations (e.g., the interpretation that there were five taps, that there were more than two taps, etc.) may live on.

Although different attributes are indeed extracted by different neural facilities at different rates (e.g., location vs. shape vs. color), and although if asked to respond to the presence of each one in isolation we would do so with different latencies, we perceive events, not a successively analyzed trickle of perceptual elements or attributes. As Efron remarks:

> There are no grounds for an a priori assumption that the specificity of our awareness of an object of perception, or an aspect of that object, gradually increases or grows following the moment of its onset from the least specific experience to some maximally specific experience. . . . We do not, when first observing an object with central vision, fleetingly experience the object as it would appear with the most peripheral vision, then as it would appear with less peripheral vision. . . . Similarly, when we shift our attention from one object of awareness to another, there is no experience of "growing" specificity of the new object of awareness – we just perceive the new object. (1967, p. 721)

Is there an "optimal time of probing"? On the plausible assumption that after a while such narratives degrade rather steadily through both fading of details and self-serving embellishment (what I ought to have said at the party tends to turn into what I did say at the party), one can justify probing "as soon as possible" after the stimulus sequence of interest. At the same time, one wants to avoid interfering with the phenomenon by a premature probe. Because perception turns imperceptibly into memory, and "immediate" interpretation turns imperceptibly into rational reconstruction, there is no single, all-context summit on which to direct one's probes. Any probe may

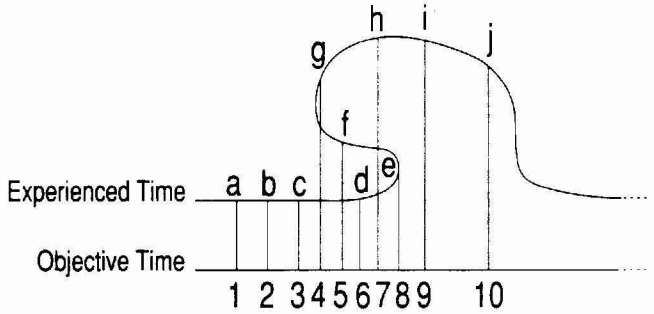

Figure 5. Superimposition of subjective and objective sequences.

elicit a narrative (or narrative fragment), and any such elicited narrative determines a "time line," a subjective sequence of events from the point of view of an observer. This time line may then be compared with other time lines, in particular with the objective sequence of events occurring in the brain of that observer. For the reasons discussed, these two time lines may not superimpose themselves in orthogonal registration. There may be order differences that induce kinks.

There is nothing metaphysically extravagant or challenging about this failure of registration (Snyder 1988). It is no more mysterious or contracausal than the realization that the individual scenes in movies are often shot out of sequence, or that when you read the sentence, "Bill arrived at the party after Sally, but Jane came earlier than either of them," you learn of Bill's arrival before you learn of Jane's earlier arrival. The space and time of the representing is one frame of reference; the space and time of what the representing represents is another. But this metaphysically innocuous fact does nevertheless ground a fundamental metaphysical category: When a portion of the world comes in this way to compose a skein of narratives, that portion of the world is an observer. That is what it is for there to be an observer in the world, a something it is like something to be.

3. The Libet controversies re-examined

3.1. Libet's experiments allegedly showing "backwards referral." Libet's experiments with direct cortical stimulation have provoked a great deal of discussion and speculation, in spite of the fact that they involved very few subjects, were inadequately controlled, and have not been replicated (Churchland 1981a; 1981b). No doubt they have attracted this unusual attention, in spite of their serious technical flaws because, according to Libet, they demonstrate "two remarkable temporal factors":

1. *There is a substantial delay before cerebral activities*, initiated by a sensory stimulus, *achieve "neuronal adequacy"* for eliciting any resulting conscious sensory experience.

2. After neuronal adequacy is achieved, the *subjective timing* of the experience *is (automatically) referred backwards in time*, utilizing a "timing signal" in the form of the initial response of cerebral cortex to the sensory stimulus (1981a, p. 182).

The "timing signal" is the primary evoked potential in the cortex 10 to 20 msec after peripheral stimulation. Libet suggests that the backwards referral is always "to" the timing signal.

Libet's model is Stalinesque: various editing processes occur prior to the moment of "neuronal adequacy," at which time a finished film is projected. How is it projected? Here Libet's account vacillates between an extreme view and a moderate view (cf. Honderich 1984):

a. *Backwards projection*: It is projected backwards in time to some Cartesian Theater where it actually runs in synch with the primary evoked potentials. (The primary evoked potentials, as "timing signals," serve rather like the slateboard used in film-making, showing the projector exactly how far back in time to project the experience.)

b. *Backwards referral*: It is projected in ordinary time, but it carries something like a postmark, reminding the viewer that these events must be understood to have occurred somewhat earlier. (In this case the primary evoked potentials serve simply as dates, which might be *represented* on the Cartesian screen by a title, "On the eve of the Battle of Waterloo" or "New York City, Summer, 1942.")

Libet's own term is "referral" and he defends it by reminding us of the "long recognized and accepted" phenomenon of spatial referral, which might suggest the moderate reading. But because he also insists that this backwards referral is "remarkable" and a challenge to the theory of "psychoneural identity," he invites the extreme interpretation.[8] And his interpretation is further supported by a passage at the close of Libet 1981:

> There is experimental evidence for the view that the subjective or mental "sphere" could indeed "fill in" spatial and temporal gaps. How else, for example, could one view that already mentioned enormous discrepancy *that is known to exist* between a subjective visual image and the configuration of neuronal activities that gives rise to the experience of the image? (p. 196)[9]

Let us consider the details. "Neuronal adequacy," which Libet estimates to require up to 500 msec of cortical activity, is determined by seeing how late, following initial stimulation, a direct cortical stimulation can interfere with the consciousness subsequently reported. Beyond that critical interval, a direct cortical stimulus would be reported by the subject to be a *subsequent* experience. (Having arrived too late for incorporation by the editing room into the "final print" of the first stimulus experience, it would appear in the next installment.) Libet's data suggest a tremendously variable editing window: "The conditioning cortical stimulus could be started more than 500 msec following the skin pulse and still modify the skin sensation, although in most cases retroactive effects were not observed with S-C intervals greater than 200 msec" (1981, p. 185). Libet is careful to define neuronal adequacy in terms of effects on subsequent unhurried verbal report: "The subject was asked to report, within a few seconds after the delivery of each pair of . . . stimuli" (1979, p. 195), and he insists that "the timing of a subjective experience must be distinguished from that of a behavioral response (such as in reaction time), which might be made before conscious awareness develops" (1979, p. 193).

This proviso permits him to defend a rival interpretation of Churchland's data. Churchland (1981a) attempted to discredit Libet's claim about the long rise time to

neuronal adequacy" for consciousness by asking subjects in an experiment to say "go" as soon as they were conscious of such a skin stimulus as those used by Libet. She reported a mean response time over 9 subjects of 358 msec, which, she argued, showed that the subjects must have achieved neuronal adequacy by the 200 msec mark at the latest (allowing time for the production of a verbal response). Libet's reply is Stalinesque: A verbal reaction can be unconsciously initiated. "There is nothing magical or uniquely informative when the motor response is a vocalization of the word 'go' instead of the more usual one of a finger tapping a button. . . . The ability to detect a stimulus and to react to it purposefully, or be psychologically influenced by it, without any reportable conscious awareness of the stimulus, is widely accepted" (Libet 1981, pp. 187–88). And to the objection, "But what did Churchland's subjects think they were doing, if not saying, as requested, just when they were conscious of the stimulus?" Libet could give the standard Stalinesque reply: They did indeed eventually become conscious of the stimulus, but by then, their verbal report had already been initiated.[10]

For this reason Libet rejects such reaction time studies as Churchland's as having "an uncertain validity as a primary criterion of a subjective experience" (1981, p. 188). He favors letting the subject take his time: "The report is made unhurriedly within a few seconds after each trial, allowing the subject to introspectively examine his evidence" (p. 188). How, then, can he deal with the rival prospect that this leisurely pace gives the Orwellian revisionist in the brain plenty of time to replace the *veridical* memories of consciousness with *false* memories? "Reporting after the trial of course requires that processes of short-term memory and recallability be operative, but this presents no difficulty for subjects with no significant defects in these abilities" (Libet, p. 188).

This begs the question against the Orwellian, who is prepared to explain a variety of effects as the result of *normal* misremembering or hallucinatory recall, in which a prior, real event in consciousness is obliterated and replaced by subsequent memories. (For related discussions, see Allport 1988, pp. 171–76; Bisiach 1988, pp. 110–12.) Has Libet let the stew cook too long, or has Churchland sampled it too soon? If Libet wants to claim a *privileged* status for his choice of probe time, he must be prepared to combat the counterarguments.

Libet comes close to pleading *nolo contendere*: "Admittedly, a report of relative timing order cannot, in itself, provide an indicator of the 'absolute' time (clock-time) of the experience: As suggested, there is no known method to achieve such an indicator" (1981, p. 188). This echoes his earlier remark that there seemed to be "no method by which one could determine the absolute timing of a subjective experience" (Libet et al. 1979, p. 193). What Libet misses, however, is the possibility that this is because there is no such moment of absolute time (cf. Harnad, unpublished; 1989).

Churchland too fails to distinguish time represented from time of representing, in her criticisms (1981a; 1981b): "The two hypotheses differ essentially on just when the respective sensations *were felt* [our emphasis]," (1981a, p. 177) and

> Even if it be supposed that the sensations arising from the simultaneous skin and LM [medial lemniscus] sensations are *felt at exactly the same time* [our emphasis], the delay in neuronal adequacy for skin stimuli may well be an artifact of the setup. (1981b, p. 494)

Suppose that all such artifacts were eliminated, and *still* the sensations are "felt at exactly the same time." Will this mean that there is a time t such that stimulus 1 is felt at t and stimulus 2 is felt at t (the anti-materialist prospect) or only that stimulus 1 and stimulus 2 are felt as (experienced as) simultaneous? Churchland doesn't discourage the inference that Libet's findings, if vindicated, would wreak havoc (as he claims) on materialism. Elsewhere, however, she correctly notes that "intriguing as temporal illusions are, there is no reason to suppose there is something preternatural about them, and certainly there is nothing which distinguishes them from spatial illusions or motion illusions as uniquely bearing the benchmark of a nonphysical origin" (1981a, p. 178). This could only be the case if temporal illusions were phenomena in which *time was misrepresented*; if the *misrepresentings* take place at the "wrong" times, something more revolutionary is afoot.

Where does this leave Libet's experiments with cortical stimulation? As an interesting but inconclusive attempt to establish something about *how the brain represents temporal order*. Primary evoked potentials may somehow serve as specific reference-points for neural representations of time, although Libet has not shown this, as Churchland's technical criticisms make clear. Alternatively, the brain keeps its representations of time more labile. We don't represent seen objects as existing on the retina, but rather as various distances in the external world. Why should the brain not also represent events as happening *when* it makes the most "ecological" sense for them to happen? When we are engaged in some act of manual dexterity, "fingertip time" should be the standard; when we are conducting an orchestra, "ear time" might capture the registration. "Primary cortical time" might be the default standard (rather like Greenwich Mean Time for the British Empire) – a matter, however, for further research.

The issue has been obscured by the fact that both proponent and critic have failed to distinguish consistently between time of representing and time represented. They talk past each other, with Libet adopting a Stalinesque position and Churchland making the Orwellian countermoves, both apparently in agreement that there is a fact of the matter about exactly when (in "absolute" time as Libet would put it) a conscious experience happens.[11]

3.2. Libet's claims about the "subjective delay" of consciousness of intention. The concept of the absolute timing of an experience is exploited in Libet's later experiments with "conscious intentions," in which he seeks to determine their absolute timing experimentally by letting the subjects, who alone have direct access (somehow) to their experiences, do *self-timing*. He asked subjects to look at a clock (a spot of light circling on an oscilloscope) *while* they experience consciously intending, and to make a judgment about the position on the clock of the spot at the onset of intention, a judgment they can later, at their leisure, *report*.

Libet is clearer than most of his critics about the importance of keeping content and vehicle distinguished:

"One should not confuse *what* is reported by the subject with *when* he may become introspectively aware of what he is reporting" (Libet 1985a, p. 559). He recognizes (p. 560), moreover, that a judgment of simultaneity need not itself be simultaneously arrived at or rendered; it might mature over a long period of time (consider, for instance, the minutes it may take the stewards at the race track to develop and then examine the photo-finish picture on which they eventually base their judgment of the winner or a dead heat).

Libet gathered data on two time series: (1) the objective series, which includes the timing of the external clock and the salient neural events: the readiness potentials (RPs) and the electromyograms (EMGs), and (2) the subjective series (as later reported), which consists of mental imagery, memories of any preplanning, and, crucially, of a single benchmark datum for each trial: a simultaneity judgment of the form: *My conscious intention (W) began simultaneously with the clock spot in position P.*

Libet seems to have wanted to approximate the elusive *acte gratuit* discussed by the existentialists (e.g., Gide 1948; Sartre 1943), the purely motiveless – and hence in some special sense "free" – choice, and as several commentators have pointed out (Breitmeyer 1985; Bridgeman 1985; Danto 1985; Jung 1985; Latto 1985) such highly unusual actions (what might be called acts of deliberate pseudorandomness) are hardly paradigms of "normal voluntary acts" (Libet 1987, p. 784). But has he in any event isolated a variety of conscious experience, however characterized, that can be absolutely timed by such an experimental design?

He claims that when conscious intentions to act (at least of his special sort) are put into registration with the brain events that actually initiate the acts, there is an offset: Consciousness of intention lags 300–500 msec behind the relevant brain events. This does look ominous to anyone committed to the principle that "our conscious decisions" *control* our bodily motions. It looks as if *we* are located in Cartesian theaters where we are shown, with a half-second tape delay, the *real* decision-making that is going on *elsewhere* (somewhere *we* aren't). We are not quite "out of the loop" (as they say in the White House), but because our access to information is thus delayed, the most we can do is intervene with last-moment "vetoes" or "triggers." One who accepts this picture might put it this way: "Downstream from (unconscious) command headquarters, I take no real initiative, am never in on the birth of a project, but do exercise a modicum of executive modulation of the formulated policies streaming through my office."

This picture is compelling but incoherent. For one thing, such a "veto" would itself have to be a "conscious decision," it seems, and hence ought to require its own 300–500 msec cerebral preparation – unless one is assuming outright Cartesian dualism (see MacKay, 1985, who makes a related point). Setting that problem aside, Libet's model, as before, is Stalinesque, and the obvious Orwellian alternative is raised by Jasper (1985), who notes that both epileptic automatisms and behaviors occurring under the effect of such drugs as scopolamine show that "brain mechanisms underlying awareness may occur without those which make possible the recall of this awareness in memory afterward." Libet concedes that this "does present a problem, but was not experimentally testable" (p. 560).[12]

Given this concession, is the task of fixing the absolute *micro*timing of consciousness ill-conceived? Neither Libet nor his critics draw that conclusion. Libet, having carefully distinguished content from vehicle – *what* is represented from *when* it is represented – nonetheless tries to draw inferences from premises about what is represented to conclusions about the absolute timing of the representing in consciousness (cf. Salter 1989). Wasserman (1985) sees the problem: "The time when the external objective spot occupies a given clock position can be determined easily, but this is not the desired result." But he then falls into the Cartesian trap: "What is needed is the time of occurrence of the internal brain-mind representation of the spot."

"*The* time of occurrence" of the internal representation? Occurrence where? There is essentially continuous representation of the spot (representing it to be in various different positions) in various different parts of the brain, starting at the retina and moving up through the visual system. The brightness of the spot is represented in some places and times, its location in others, and its motion in still others. As the external spot moves, all these representations change, in an asynchronous and spatially distributed way. Where does "it all come together at an instant in consciousness"? Nowhere. Wasserman correctly points out that the task of determining where the spot was at some time in the subjective sequence is itself a voluntary task, and initiating it presumably takes some time. This is difficult not only because it is in competition with other concurrent projects (as stressed by Stamm 1985, p. 554), but also because it is unnatural – a conscious judgment of temporality of a sort that does not normally play a role in behavior control, and hence has no natural meaning in the sequence. The process of interpretation that eventually fixes the judgment of subjective simultaneity is itself an artifact of the experimental situation, and *changes the task*, therefore telling us nothing of interest about the actual timing of normal representational vehicles anywhere in the brain.

Stamm likens the situation to Heisenbergian uncertainty: "Self-monitoring of an internal process interferes with that process, so that its precise measurement is impossible" (p. 554). This observation betrays a commitment to the mistaken idea that *there is* an absolute time of intersection, "precise measurement" of which, alas, is impossible for Heisenbergian reasons (see also Harnad 1989). This could only make sense on the assumption that there is a particular privileged place where the intersection matters.

The all too natural vision that we must discard is the following: Somewhere deep in the brain an act-initiation begins; it starts out as an unconscious intention, and slowly makes it way to the theater, picking up clarity and power as it goes, and then, at an instant, t, it bursts on stage, where a parade of visual spot-representations are marching past, having made their way slowly from the retina, getting clothed with brightness and location as they moved. The audience or *I* is given the task of saying which spot-representation was "on stage" exactly when the conscious intention made its bow. Once identified, this spot's time of departure from the retina can be

calculated, as well as the distance to the theater and the transmission velocity. That way we can determine the exact moment at which the conscious intention occurred in the Cartesian Theater.

Some have thought that although that particular vision is incoherent, one does not need to give up the idea of absolute timing of experiences. There is an alternative family of models for the onset of consciousness that avoids the preposterousness of the Cartesian-centered brain. Couldn't consciousness be a matter not of arrival at a point but rather a matter of a representation exceeding some threshold of activation over the whole cortex or large parts thereof? 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 criterial level.

The idea that content becomes conscious not by entering a subsystem, but by the brain's undergoing a state change of one sort or another has much to recommend it (see, e.g., Crick & Koch 1990; Kinsbourne 1988; Neumann 1990). Moreover the simultaneities and sequences of such mode-shifts can presumably be measured by outside observers, providing, in principle, a unique and determinate sequence of contents attaining the special mode. But this is still the Cartesian Theater if it is claimed that the real ("absolute") timing of such mode shifts is definitive of subjective sequence. The imagery is different, but the implications are the same. Conferring the special property that makes for consciousness at an instant is only half the problem; discriminating that the property has been conferred at that time is the other, and although scientific observers with their instruments may be able to do this with microsecond accuracy, how is the brain to do this? We human beings do make judgments about simultaneity and sequence among elements of our own experience, some of which we express, so at some point or points in our brains the corner must be turned from the actual timing of representations to the representation of timing. This is a process that takes effort in one way or another (Gallistel 1990), and wherever and whenever these discriminations are made, thereafter the temporal properties of the representations embodying those judgments are not constitutive of their content.

Suppose that a succession of widely spread activation states, with different contents, sweeps over the cortex. The actual, objectively measured simultaneities and sequences in this broad field are of no functional relevance *unless they can also be accurately detected by mechanisms in the brain*. What would make *this* sequence the stream of consciousness if the brain could not discern the sequence? 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.

3.3. Grey Walter's experiment: A better demonstration of the central contention of the Multiple Drafts model. It was noted above that Libet's experiment created an artificial and difficult judgmental task that robbed the results of the hoped-for significance. This can be brought out more clearly by comparing it to a similar experiment by Grey Walter (1963), with patients in whose motor cortex he had implanted electrodes. He wanted to test the hypothesis that certain burst of recorded activity were the initiators of intentional actions, so he arranged for each patient to look at slides from a carousel projector. The patient could advance the carousel at will, by pressing the button on the controller. (Note the similarity to Libet's experiment: This was a "free" decision, timed only by an endogenous rise in boredom, or curiosity about the next slide, or distraction, or whatever.) Unbeknownst to the patient, however, the controller button was a dummy, not attached to the slide projector at all. What actually advanced the slides was the amplified signal from the electrode implanted in the patient's motor cortex.

One might suppose that the patients would notice nothing out of the ordinary, but in fact they were startled by the effect, because it seemed to them as if the slide projector was anticipating their decisions. They reported that just as they were "about to" push the button, but before they had actually decided to do so, the projector would advance the slide – and they would find themselves pressing the button with the worry that it was going to advance the slide twice! The effect was strong, according to Grey Walter's account, but apparently he never performed the dictated followup experiment: introducing a variable delay element to see how large a delay had to be incorporated into the triggering to eliminate the "precognitive carousel" effect.

An important difference between Grey Walter's and Libet's designs is that the judgment of temporal order that leads to surprise in Grey Walter's experiment is part of a normal task of behavior monitoring. In this regard it is like the temporal order judgments by which our brains distinguish moving left-to-right from moving right-to-left, rather than "deliberate, conscious" order judgments. The brain in this case has set itself to "expect" visual feedback on the successful execution of its project of advancing the carousel, and the feedback arrives earlier than expected, triggering an alarm. This could show us something important about the actual timing of content vehicles and their attendant processes in the brain, but it would not, contrary to first appearances, show us something about the "absolute timing of the conscious decision to change the slide."

Suppose, for instance, that an extension of Grey Walter's experiment showed that a delay as long as 300 msec (as implied by Libet) had to be incorporated into the implementation of the act in order to eliminate the subjective sense of precognitive slide-switching. What such a delay would in fact show would be that expectations set up by a decision to change the slide are tuned to expect visual feedback 300 msec later, and to report back with alarm under other conditions. The fact that the alarm eventually gets interpreted in the subjective sequence as a perception of misordered events (change before button push) shows nothing about *when* in real time the consciousness of the decision to press the button first occurred. The sense the subjects reported of not quite having had time to "veto" the initiated button push when they "saw the slide was already changing" is a natural interpretation for the brain to settle on (eventually) of the various contents made available at various times for incorporation into the narrative. Was this sense already there

at the first moment of consciousness of intention (in which case the effect requires a long delay to "show time" and is Stalinesque) or was it a retrospective reinterpretation of an otherwise confusing *fait accompli* (in which case it is Orwellian)? This question should no longer seem to demand an answer.

4. Conclusion

The Multiple Drafts model has many other implications for scientific theories of consciousness (Dennett 1991b), but our main conclusion in this target article is restricted to temporal properties of experience: The representation of sequence in the stream of consciousness is a product of the brain's interpretative processes, not a direct reflection of the sequence of events making up those processes. Indeed, as Jackendoff has pointed out to us, what we are arguing for in this essay is a straightforward extension to the experience of time of the common wisdom about the experience of space; the representation of space in the brain does not always use space-in-the-brain to represent space, and the representation of time in the brain does not always use time-in-the-brain. It may be objected that the arguments presented here are powerless to overturn the still obvious truth that our experiences of events occur in the very same order that we experience them to occur. If someone thinks the thought, "One, two, three, four, five," his thinking "one" occurs before his thinking "two" and so forth. The example does illustrate a thesis that is true in general and does indeed seem unexceptioned, so long as we restrict our attention to psychological phenomena of "ordinary," macroscopic duration. But the experiments we selected for discussion are concerned with events that were constricted by unusually narrow timeframes of a few hundred milliseconds. At this scale, we have argued, the standard presumption breaks down.

It might be supposed, then, that we are dealing only with special cases. These limiting cases may interestingly reveal how the brain deals with informational overload, but, one might suggest, they are unrepresentative of the brain's more usual manner of functioning. The contrary is the case, however, as might be anticipated, in view of the brain's well-known propensity for applying a limited number of basic mechanisms across a wide range of situations. The processes of editorial revision that are dramatically revealed in the time-pressured cases continue indefinitely as the brain responds to the continued demands of cognition and control. For instance, as time passes after an event has occurred, that event may be recalled to episodic memory, but to an ever more limited extent. After some days, an occurrence that may have unrolled over minutes or more is remembered within as restricted a time frame as those we have been discussing. Such memories present not as randomly blurry or depleted versions but as internally coherent, simplified renderings of what are taken to be the most important elements. Temporal succession is typically an early victim of this reorganization of the event, sacrificed in favor of (apparently) more useful information (as instanced in the phi phenomenon).

We perceive – and remember – perceptual events, not a successively analyzed trickle of perceptual elements or attributes locked into succession as if pinned into place on a continuous film. Different attributes of events are indeed extracted by different neural facilities at different rates, (e.g., location vs. shape vs. color) and people, if asked to respond to the presence of each one in isolation, would do so with different latencies, depending on which it was, and on other well-explored factors. The relative timing of inputs plays a necessary role in determining the information or content of experience, but it is not obligatorily tied to any stage or point of time during central processing. How soon we can respond to one in isolation, and how soon to the other, does not exactly indicate what will be the temporal relationship of the two in percepts that incorporate them both.

There is nothing theoretically amiss with the goal of acquiring precise timing information on the mental operations or informational transactions in the brain (Wasserman & Kong 1979). It is indeed crucial to developing a good theory of the brain's control functions to learn exactly when and where various informational streams converge, when "inferences" and "matches" and "bindings" occur. But these temporal and spatial details do not tell us directly about the contents of consciousness. The temporal sequence *in consciousness* is, within the limits of whatever temporal control window bounds our investigation, purely a matter of the content represented, not the timing of the representing.

ACKNOWLEDGMENTS

The original draft of this essay was written while the authors were supported by the Rockefeller Foundation as Scholars in Residence at the Bellagio Study Center, Villa Serbelloni, Bellagio, Italy, April, 1990. We are grateful to Kathleen Akins, Peter Bieri, Edoardo Bisiach, William Calvin, Patricia Churchland, Robert Efron, Stevan Harnad, Douglas Hofstadter, Tony Marcel, Odmar Neumann, Jay Rosenberg, and David Rosenthal for comments on subsequent drafts.

NOTES

1. A philosophical exception is Vendler (1972; 1984) who attempts to salvage Cartesian dualism. A scientific exception is Eccles (e.g., Popper & Eccles 1977).

2. What about the prospect of a solitary Robinson Crusoe scientist who performs all these experiments wordlessly on himself? Would the anomalies be apparent to this lone observer? What about reconstructing these experiments with languageless animals? Would we be inclined to interpret the results in the same way? Would we be justified? These are good questions, but their answers are complicated, and we must reserve them for another occasion.

3. Such a "postmark" can be in principle be added to a vehicle of content at any stage of its journey; if all materials arriving at a particular location come from the same place, by the same route at the same speed, their "departure time" from the original destination can be retroactively stamped on them, by simply subtracting a constant from their arrival time at the way station. This is an engineering possibility that is probably used by the brain for making certain automatic adjustments for standard travel times.

4. "The essence of much of the research that has been carried out in the field of sensory coding can be distilled into a single, especially important idea – any candidate code can represent any perceptual dimension; there is no need for an isomorphic relation between the neural and psychophysical data. Space can represent time, time can represent space, place can represent quality, and certainly, nonlinear neural functions can represent linear or nonlinear psychophysical functions equally well" (Uttal 1979). This is a widely acknowledged idea,

but, as we will show, some theorists (mis-)understand it by tacitly reintroducing the unnecessary "isomorphism" in a dimly imagined subsequent translation or "projection" in consciousness.

5. Cf. Pylyshyn 1979: "No one . . . is disposed to speak *literally* of such physical properties of a mental event as its color, size, mass, and so on – though we *do* speak of them as *representing* (or having the experiential content of) such properties. For instance, no one would not properly say of a thought (or image) that it was large or red, but only that it was a thought *about* something large or red (or that it was an image *of* something large or red). . . . It ought to strike one as curious, therefore, that we speak so freely of the *duration* of a mental event."

6. P. S. Churchland (1981a, p. 172) notes a difference between "masking in the usual sense" and "blanking in short term memory," which perhaps is an allusion to these two possibilities, but does not consider how one might distinguish between them.

7. Consider the medio-temporal region of cortex (MT), which responds to motion (and apparent motion). Suppose then that some activity in MT *is* the brain's concluding that there was intervening motion. There is no further question, on the Multiple Drafts model, of whether this is a pre-experiential or post-experiential conclusion. It would be a mistake to ask, in other words, whether this activity in MT was a "reaction to a conscious experience" (by the Orwellian historian) as opposed to a "decision to represent motion" (by the Stalinesque editor).

8. See also his dismissal of MacKay's suggestion of a more moderate reading (Libet 1981, p. 195; 1985b, p. 568).

9. Libet's final summation in 1981, on the other hand, was inconclusive: "My own view . . . has been that the temporal discrepancy creates relative difficulties for identity theory, but that these are not insurmountable" (p. 196). Presumably they would be undeniably insurmountable on the backwards *projection* interpretation, and Libet later (1985b, p. 569) describes these difficulties in a way that seems to require the milder reading: "Although the delay-and-antedating hypothesis does not separate the actual time of the experience from its time of neuronal production, it does eliminate the necessity for simultaneity between the *subjective timing* of the experience and the actual clock-time of the experience." Perhaps Eccles's enthusiastic support for a radical, dualistic interpretation of the findings has misdirected the attention of Libet (and his critics) from the mild thesis he sometimes defends.

10. In an earlier paper, Libet conceded the possibility of Orwellian processes and supposed there might be a significant difference between unconscious mental events and conscious-but-ephemeral mental events: "There may well be an immediate but ephemeral kind of experience of awareness which is not retained for recall at conscious levels of experience. If such experiences exist, however, their content would have direct significance only in later unconscious mental processes, although, like other unconscious experiences, they might play an indirect role in later conscious ones" (1965, p. 78).

11. Harnad (1989) sees an insoluble problem of measurement, but denies our contention that there is no fact of the matter: "Introspection can only tell us when an event *seemed* to occur, or which of two events *seemed* to occur first. There is no independent way of confirming that the real timing was indeed as it seemed. Incommensurability is a methodological problem, not a metaphysical one." So Harnad asserts what we deny: that among the real timings of events in the brain is a "real timing" of events *in consciousness*.

12. In a later response to a similar suggestion of Hoffman and Kravitz (1987) Libet asks the rhetorical question, "Are we to accept the primary evidence of the subjects' introspective report (as I do), or are we going to insist that the subject had a conscious experience which he himself does not report and would even deny having had?" (1987, p. 784). This is another expression of Libet's a priori preference for a Stalinesque position.