RUSSELL ON MNEMIC CAUSATION

SVEN BERNECKER
University of Munich

Abstract

According to the standard view, the causal process connecting a past representation and its subsequent recall involves intermediary memory traces. Yet Bertrand Russell and Ludwig Wittgenstein held that since the physiological evidence for memory traces isn't quite conclusive, it is prudent to come up with an account of memory causation-referred to as mnemic causation—that manages without the stipulation of memory traces. Given mnemic causation, a past representation is directly causally active over a temporal distance. I argue that the stipulation of memory traces is indeed indispensable for analyzing memory causation.

A claim to remember a past event implies not merely that the rememberer represented (or experienced) such an event, but that her present memory is in some way *due to*, that it came about *because of*, a cognitive and sensory state she had at the time she represented (or experienced) the event. Remembering that *P* implies that one's recall that *P* is causally derived from one's previous representation of *P*.[1]

According to the standard view, the causal process connecting a past representation and its subsequent recall involves intermediary memory traces (or engrams). The stipulation of memory traces is motivated by the contention that between any two diachronic mental events there have to be a series of intermediary events, each of which causes the next, and each of which is temporally contiguous to the next. However, in *The Analysis of Mind*, Bertrand Russell argued that since the physiological evidence for memory traces is not yet quite conclusive, it is prudent to come up with an account of

© *Principia*, 5(1–2) (2001), pp. 149–85. Published by NEL — Epistemology and Logic Research Group, Federal University of Santa Catarina (UFSC), Brazil.

memory causation that manages without the stipulation of memory traces. Russell's name for this notion of memory causation is 'mnemic causation'. According to the theory of mnemic causation, a past experience is directly causally active over a temporal distance; past experiences are proximal causes of states of recalling. The aim of this paper is to refute the theory of mnemic causation and to establish that the stipulation of memory traces is indispensable for the analysis of memory.

Sections 1 and 2 motivate the standard view that causality implies contiguity and explicate the notion of a memory trace. Section 3 is a discussion of the verifiability of memory traces. Section 4 provides a detailed account of Russell's theory of mnemic causation. The concept of mnemic causation is defended against the widespread objection that direct causal action at a distance in time is impossible because a cause cannot operate when it has ceased to exist. Given that Russell's notion of mnemic causation does not face any crucial difficulties and given that there is no conclusive empirical evidence for the existence of memory traces, how should we determine whether or not to stipulate memory traces? My thesis is that the apparent tie between the theory of mnemic causation and the theory of contiguous causation is resolved as soon as we treat memory traces as theoretical constructs. For when traces are treated as theoretical constructs, it can be shown that the concept of a trace is indispensable to explain certain features of our intuitive notion of memory. This is done in sections 6 and 7.

1. Contiguous Causation

Saying that a representation of *P* qualifies as a memory if and only if it is causally connected to a previous representation of *P*, raises the question of how one should conceive of the causal connection. David Hume is the dominant philosopher of cause and effect. He notoriously maintained that two events are related as cause and effect only if they meet each of three individually necessary and jointly sufficient conditions: (1) priority of cause to effect, (2) contiguity in space and time, and (3) constant conjunction or necessary connection (1978, pp. 73–8). Each one of these conditions has spawned an enormous

and still ongoing debate. Fortunately, we need to concern ourselves only with two of the three conditions. We can abstract from the first condition that a cause must precede its effect in time. For in the case of *memory* causation, it is indisputable that the cause can neither be simultaneous with its effect (concurrent causation), nor temporally posterior to its effect (backward causation).

Regarding the condition of temporal and spacial contiguity, Hume declared in the *Treatise of Human Nature*, "nothing can operate in a time or place, which is ever so little remov'd from those of its existence".[2] The reason Hume was led to stipulate that contiguity is a necessary condition of causation is that if cause and effect were not contiguous, some factor could intervene and prevent the effect, even though the cause had occurred. As Hume himself noticed, we are often not aware of the continuous causal paths connecting cause and effect. A switch on the wall is some distance from the electric light overhead that it controls; pushing a button on an alarm clock makes it ring seven hours later. Where contiguity appears to be lacking, Hume held that we find, upon closer examination, that they *are* connected by a chain of causes such that the effect is finally caused by an event that is contiguous with it.

> Tho' distant objects may sometimes seem productive of each other, they are commonly found upon examination to be link'd by a chain of causes, which are contiguous among themselves, and to the distant objects; and when in any particular instance we cannot discover this connexion, we still presume it to exist (1978, p. 75).

In other words, by making a distinction between remote and proximate causes, we may say that the remote cause is connected with the effect through a chain of causes, the last one being the proximate cause. And that the proximate cause is that event that is contiguous with and produces the effect. Hume concluded that contiguity is 'essential' to causality. The causal path has no spatial or temporal gaps or breaks.

Today, the prevailing view is still that causality implies contiguity. Ernest Nagel, for example, writes:

> [T]he [causal] relation has a temporal character, in the sense that the event said to be the cause precedes the effect and is also 'con-

> tiguous' with the latter. In consequence, when events separated by a temporal interval are said to be causally related, they are also assumed to be connected by a series of temporally adjacent and causally related events (1961, p. 74).

And Alfred Ayer declares, "[i]t is fairly generally assumed that in the cases where the cause can be represented as an event which precedes the effect, the two events must be temporally contiguous" (1972, p. 135).

Like any philosophical position, the thesis of contiguous causation is open to criticism. An important objection lodged against this thesis stems from quantum mechanics. Quantum mechanics seems to permit non-instantaneous action at a distance where no energy exists in the space across which the action occurs. But setting aside physics, of which I know nothing, there is another interesting objection to the thesis of contiguous causation. In his early paper, "On the Notion of Cause," Russell developed the *simultaneity paradox* which is supposed to show that a cause cannot be temporally contiguous with its effect.

1.1. Simultaneity Paradox

The simulaneity paradox has the form of a reductio: Russell assumes that cause and effect are temporally contiguous and shows that when this thesis is conjoined with the idea of necessary connection, it entails that cause and effect are contemporaneous. The argument conceives of cause and effect as events that take time and are divisible into atomic units. Consider, for example, the breaking of a cup on the kitchen floor. The cause begins with the knocking over by one's elbow, and encompasses the downward hurling until the impact. The effect begins with the first impact on the floor, and ends with distribution of the pieces of china on the floor. Russell maintains that a real cause takes place only in the instant directly adjoined to the instant in which the effect begins. And the real effect takes place only in the instant right after the final unit of the cause-event has ceased to exist. For, if there is the slightest interval between the cause and effect, "something may happen during the interval which prevents the

expected result," even though 'the cause' had occurred.[3] In the example at hand, if the cause were some event prior to the moment of impact on the kitchen floor (e.g., the knocking over), then any number of occurrences may intervene (e.g., the sudden disappearance of gravity), altering the normal course of events so that the cup does not break. But if the effect may not have happened, then the purported cause is not a necessary condition for the occurrence of the effect and hence not a genuine cause. Hence, if causation is analyzed in terms of necessary conditions, cause and effect must be perfectly simultaneous; there may be no temporal interval between cause and effect.

Prima facie, it is easy to rebut Russell's simultaneity paradox and to hold on to the idea of contiguous causation. All one has to do is to reject the Humean analysis of causation in terms of necessary conditions. When the thesis of contiguous causation is combined with, say, a probabilistic notion of causation,[4] the simultaneity paradox evaporates. However, this strategy misjudges the point of Russell's simultaneity paradox. In my view, the point of the simultaneity paradox is not to refute the doctrine of contiguous causation but to point out that contiguous causation presupposes a contentious claim about the nature of time, namely that time is *discrete*.[5] To say that time is discrete is tantamount to saying that time has a granular structure, with there being a smallest quantum of time. To see that contiguous causation assumes the discreteness of time, suppose that time were infinitely divisible, like real numbers are. Assuming the continuity of time, there would be a time interval separating any two causally related events which are not simultaneous. And, during this interval, something could happen which prevents the effect from occurring, although the cause had occurred. Hence, arguing for the thesis of contiguous causation requires, among other things, an argument for the discreteness of time. The formulation of such an argument would, of course, go beyond of the scope of this paper.

2. What are Memory Traces?

Notwithstanding the difficulties surrounding the idea of contiguous causation, this idea forms the basis of most philosophical accounts of

memory. It underlies all those theories of memory which explicate the causal process connecting a past experience and a subsequent recall by means of a continuous memory trace (or a sequence of traces, respectively). The trace hypothesis states that between any two diachronic mental events there is a series of intermediary events, each of which causes the next, and each of which is temporally contiguous to the next. A rigorous definition of the trace hypothesis would tell us exactly how short the time span between mental event *A* and mental event *B* has to be, for *A* to act as the direct cause of *B*. (In the case of short-term visual memory, the temporal distance between the original experience and the remembering may be as short as 0.25 seconds.)

The stipulation of memory traces is backed by common sense. How else can past representations (or experiences) act at a temporal distance, if not through a continous trace (or a series of traces)? How can there be direct causes remote in space and time? If contiguity were not a necessary component of memory, remembering, it seems, would have to rely on a magical process bearing some resemblance to telepathy and clairvoyance. Furthermore, without causal continuity, a past thought or event would somehow have to track one's spatiotemporal path, to ensure that it could, at any time, become causally active as one moves around. Such long-distance tracking of past thoughts or events seems unlikely to be 'direct' in any intuitive sense.

Memory traces are designed to account both for the *propagation* of information and for the *production* of states of recall. The stipulation of memory traces allows us to understand how past experiences can exert causal influence long after they have ceased to exist. Furthermore, by means of postulating traces, we can explain the transmission of information through time. Corresponding to the two aspects of causal processes—production and propagation—there are two distinct aspects of the notion of a memory trace: a *mental* and a *physical* aspect. Insofar as memory traces produce states of recall, they may be purely physical states. To account for the production-aspect of memory causation, it suffices to conceive of traces in purely physical terms. From a physicalist point of view, memory traces are structural modifications at synapses (i.e., the area where the axion of

one neuron connects with the dentrite of another neuron) that affect the ease with which neurons in a neural network can activate each other.[6]

2.1. Dispositional Belief and Subdoxastic State

Insofar as memory traces communicate information, they have to be capable of bearing content, and hence, they have to be mental states. What kind of mental states are memory traces? The representational nature of traces varies depending on the kinds of memory they give rise to. In this context, we need to differentiate between two kinds of memories: object-, property-, and event memory, on the one hand, and fact memory, on the other. An expression of an object memory is, "I remember the dog Fido"; an expression of a property memory is, "I remember Fido's floppy ears"; an expression of an event memory is, "I remember Fido biting the mailman"; and an expression of fact memory is, "I remember that Fido bit the mailman".[7]

Fact memory of previous fact awareness presupposes the possession of the relevant concept. If I didn't have the concept of a dog and of a sofa, I could not believe that Fido is on the sofa. And if I could not *believe* that Fido is on the sofa, I could not *remember* it either. Attributions of beliefs and fact memories are limited by the concepts possessed by the subject. Therefore, it is reasonable to assume that the traces constitutive of fact memory of previous fact awareness are *dispositional beliefs*.[8]

Object-, property-, and event memory, however, do not require that the rememberer possesses the concepts necessary for expressing the memories in question. I could, for example, remember having seen Fido on the sofa, even if I didn't possess the concept of a dog and of a sofa. Due to my limited conceptual abilities, I wouldn't know what it is that I remember, but I would remember the event nevertheless. My non-conceptual memory would allow me to discriminate this event from others, even if not by thinking or speaking of it as involving a dog and a sofa. Memory traces constitutive of object-, property, and event memories may contain non-conceptual information. Unconscious states capable of transmitting non-conceptual content are commonly called '*subdoxastic states*'.[9] In sum, traces of fact-memory

of previous fact awareness are dispositional beliefs, while traces of object-, property-, and event memory are subdoxastic states.[10]

2.2. Connectionism and Language of Thought

It a basic assumption of cognitive science that there are three levels of explanation for any intentional process. First, there is the intentional level of everyday psychology, in which we talk about people's beliefs, memories, desires and other such intentional states. Second, there is the computational level which explains how intentional states are realized by means of computational operations. These operations could be instantiated by any physical system with a sufficient degree of complexity, whether it be a human brain, a von Neumann computer, or some other cognitive system. Finally there is the level of physical implementation or realization. The three-floor model of the mind is of course inspired by the distinction between the hardware, the software, and the system interface of a computer.

So far I have only talked about the intentional and the physical level of memory traces. But what about the computational aspect? The computational account of traces depends on whether one endorses the *classical* or the *connectionist* approach. The classical approach holds that mental representations are symbolic structures that have semantically evaluable constituents and mental processes are rule-governed manipulations of them. This position lends itself particularily well to account for memorial metarepresentations of past contents (e.g., I remember that I believed that *P*). On the assumption that thoughts are syntactically structured, we can conceive of mental metarepresentation in analogy to linguistic metarepresentation. Memories of one past mental states are to be conceived of in analogy to direct quotation.

Connectionism has it that mental representations are realized by patterns of activation in a network of model neurons and mental processes consist of the spreading activation of such patterns. Given this model, information is not stored by a formulae in an internal code with a specific location. Rather information is encoded through a change in the strengths of connections between nodes. What is stored in memory is a set of changes in the instructions neurons send

each other, affecting what patterns of activity can be constructed from given inputs. Since many items of information can be represented over the same set of neurons and connections, connectionist networks store efficiently. Remembering occurs when an input 'travels' through an already established activation pattern. Each memory is a product of the activity of the total neural network. As David Rumelhart and Donald Norman say, "[i]nformation is not stored anywhere in particular. Rather it is stored everywhere" (1981, p. 3). An immediate consequence of connectionism is that memories are deeply sensitive to context. Context produces slight differences in activity patterns and the corresponding memory may be subtly different on different occasions (though these difference may not be noticed by the rememberer).

An important argument in favor of connectionist models of memory stems from neurology. It is a common fact that brain damage may not result in a sudden loss of certain kinds of information but that the performance of the memory system becomes slowly worse. Psychologists refer to this phenomenon as 'graceful degradation'. Connectionist models also exhibit smooth degradation in the face of 'lesions', i.e., removal of processing nodes and alteration of connection weights. This suggests that memory traces are distributed across many different brain cells, rather than located in one specific cluster of cells (Rose 1992).

Figure 1

One of the fundamental characteristics of memory is its reconstructive nature. Our recall of events and thoughts is frequently not literal; rather, we reconstruct memories of past events that contain inferences and these inferences, may be indistinguishable from 'real' happenings. Assuming the connectionist model, it is easy to explain the reconstructive nature of our memory. Since memories are encoded in the connection weights, memories are shared over the same hardware. Retrieval is more a matter of reconstructing information

than going to a discrete location to find it. It is therefore not surprising that connectionist networks are good in modeling *pattern completion*. Simple examples of pattern completion are when we identify letters and words even though they are presented only for a split second or they are presented incompletely. Figure 1 shows a word containing three ambiguous characters that each constrain the identity of the others (cited by Baddeley 1990, p. 366). Connectionist networks have the astonishing capacity to take such an ambiguous stimulus and to quickly give an unambiguous response. As connectionist computers we don't need to perceive every letter in a string of words to be able to read it. To xllxstxatx, I cxn rxplxce xvexy txirx lextex of x sextexce xitx an x, anx yox stxll xan xanxge xo rxad xt—ix wixh sxme xificxltx (Anderson 1995, p. 62).

A feature of our memory related to pattern completion is *content addressability*. Content addressability (a term from computer science) means starting retrieval with part of the content of the to-be-remembered material, which provides an 'address' to the place in memory where identical or similar material is located. For example, if I tell you that I am trying to remember the name of J. F. Kennedy's blond girlfriend who later married Arthur Miller, you can, no doubt, provide a great deal of associated information in addition to the name 'Marilyn Monroe'. You may, for example, remember the missile crisis, Kennedy's assassination, and the movie "Some Like It Hot". This way of getting into memory, by matching the current contents of experience to similar contents in memory and then retrieving associated information, is our primary way of remembering. Given connectionism, the content addressability of memory is due to the fact that the incoming pattern of activation has matching parts to a previous pattern and that this is sufficient to reactivate other parts of the pattern. Content addressable memory is characteristic of humans, but hard to achieve in classical architectures, where items are typically accessed on the basis of knowing in what register they were stored.

Finally, a feature of human memory which connectionism is particularly apt in accounting for is *cross-talk*. Cross-talk occurs when, in an attempt to activate a particular memory trace, similarity to another trace leads to the alternative trace being reinstated rather than the 'intended' trace. An example of cross-talk is when someone goes

upstairs to her bedroom with the intention of changing out of her T-shirt into a pullover but instead undresses completely (cf. Reason and Mycielska 1982). Such errors are expected to emerge from a connectionist model of memory where a net for a planned action overlaps with the net for a very familiar yet different sequence. The activation of the less familiar trace might trigger the more familiar and so more strongly weighted trace.

3. The Verifiability of Memory Traces

After having explained the notion of a memory trace we can turn to Russell's critique of the trace hypothesis. Two of the fifteen lectures that constitute Russell's *Analysis of Mind* deal with the issue of memory traces. Both of these lectures are, in large part, a response to the work of the zoologist Richard Semon, whom Russell regarded as "the best writer on mnemic phenomena" (1995, p. 83).

Semon was a passionate advocate of the trace hypothesis. Although he admitted that science had not progressed enough to say what memory traces are and how they work, Semon was deeply convinced that traces consist in something organic, in "a material alteration" (1909, pp. 138–9). Commenting on this aspect of Semon's position, Russell writes:

> Concerning the nature of an engram, Semon confesses that at present it is impossible to say more than that it must consist in some material alteration in the body of the organism. [...] It is, in fact, hypothetical, invoked for theoretical uses, and not an outcome of direct observation. No doubt physiology, especially the disturbances of memory through lesions in the brain, affords grounds for this hypothesis; nevertheless it does remain a hypothesis (1995, p. 85).

This passage illustrates Russell's acute awareness of the fact that, in his time, the trace hypothesis was empirically underdetermined. Russell holds that the empirical evidence for traces is "not quite conclusive" (1995, pp. 86, 92), but he regards it to be "quite possible" that some day physiological memory traces will be discovered. Given "the present state of physiology", Russell writes, "the introduction of the

engram does not serve to simplify the account of mnemic phenomena", and it is therefore prudent to settle for a less speculative account of memory causation, an account formulated "in terms, wholly, of observable facts".[11]

Russell's point is that if traces exist, they must be shown to exist and not simply postulated to support the account of memory causation. As long as traces cannot be proven empirically, one should refrain from assuming their existence. And this is exactly what Russell did in *The Analysis of Mind*. Because of the inconclusiveness of the physiological evidence in favor of memory traces, Russell sees the necessity to develop a notion of memory causation that manages without the stipulation of memory traces. Russell's name for this notion of memory causation is *mnemic causation*. Before examining the concept of mnemic causation (cf. section 4), I want to take a closer look at the contention that the physiological evidence for memory traces fails to be conclusive.

Russell is certainly right in maintaining that brain-injured patients tend to have deficits that are rarely sufficiently pure, or specific enough, to draw any interesting conclusions regarding the locations of the memory system in the brain. Granted that the disturbances of memory through brain damage fail to provide conclusive evidence for the trace hypothesis, what would constitute conclusive evidence?

3.1. Neurobiological Evidence

Russell presumably holds that for neurobiology to *prove* the trace hypothesis, it has to establish the identity between memories, on one hand, and structural modifications of the synapses, on the other. A neurosurgeon could then, by artificially structuring synapses, bring about a certain memory; and by removing certain synapses she could erase that memory. Russell regards it as "quite possible" that some day neurobiology will come up with one-to-one correlations between memory states and brain states. Yet this day hasn't come yet, and it is questionable whether it will come soon.

In recent years, a number of fascinating non-invasive diagnostic methods for studying brain operations have been developed. These techniques make it possible to 'observe' in vivo how the human brain

subserves cognition. CAT scans, for example, detect activity by the enzyme choline-acetyltransferase, thereby allowing one to trace the links between cells. PET scans (positron emission tomography scans) display radioactive emitters tagged to glucose molecules carried in the bloodstream to the brain. The radioactive tag may be put on a precursor chemical that will go on to form a neurotransmitter; or a drug may be tagged and followed as it courses throughout the brain. In both cases, brain operations and chemical transformations can be monitored. MEG scans (magnetoencephalography scans) use multiple rapidly oscillating magnetic field gradients. Hydrogen atoms resonate in the water molecules of living tissue, and this resonance can be detected by radiowaves and a large electromagnet. It allows one to calculate the density of brain tissue.

Despite all these fancy methods for studying brain processes, it hasn't yet been possible to establish a one-to-one correspondence between memories and brain states. The problem is that the data delivered by these diagnostic methods are not as fine-grained as the contents of our memories. None of the above-mentioned diagnostic methods can, for example, account for the difference between the memory that triangles have three sides and the memory that triangles have three angles. Another problem is that memory encoding structures of the brain can change in response to further experiences. In sum, given the current development of neurobiological techniques, it is impossible to rule out the possibility that two memories (had by the same subject) which differ in content supervene on the same brain process and thus are realized by the same trace (cf. Mayes 2001, p. 191). As long as such cases cannot be ruled out, Russell's claim that empirical evidence for the trace hypothesis is inconclusive, stays valid.

Direct brain stimulation doesn't suffer from the defect of being insufficiently specific. During the 1940's the Canadian neurosurgeon Wilder Penfield carried out brain operations on epileptic patients in order to relieve their intractable seizures. Patients were first fully anaesthetized; the appropriate area of the skull was then removed, and the brain exposed. Consciousness was then restored, with only a local anaesthetic being maintained. During the operation, it was necessary to stimulate the surface of the brain with an electrode. Pen-

field noted that electrical stimulation in the temporal lobes resulted in patients having memory flashbacks. One patient said, "I just heard one of my children speaking [...] it was Frank, and I could hear the neighborhood noises," and another, "[s]omething brings back a memory. I can see the Seven-Up bottling company" (1958). In general, these recollections were vivid, detailed, and concerned with seemingly insignificant past events.

Although suggestive, Penfield's experiments fail to provide conclusive evidence for the hypothesis of physical memory traces. First of all, of the 520 patients who received electrical stimulation in the temporal lobes, only 40 reported having memory flashbacks. Secondly, Penfield failed to test whether repeated stimulation of the same area of the lobes gave rise to the same type of memory. Thirdly, subsequent studies have shown that such memory flashbacks occur only when the limbic structures (generally believed to be essential for emotional experiences) are activated (cf. Gloor et al. 1982).

In sum, until neurobiology has progressed enough to assign a brain state to every memory state, it is pure speculation that there is a one-to-one correspondence between brain states and memory states. (Given connectionism, it is exceedingly difficult to establish a one-to-one correspondence between brain states and memory states. For according to connectionism, information is stored in the relationship between neurons and each neuron participates in the encoding of many different memories.)

3.2. Introspective Evidence

Granted that the third-person data of today's neurobiology present inconclusive evidence for the trace hypothesis, maybe the existence of memory traces can be proven by means of introspective data from a first-person perspective. However, we don't have introspective access to our brain states. If anything, it is the intentional contents of memory traces that can be detected via introspection. But is it really the case that we have direct introspective access to the contents of our memory traces? Are memory traces transparent to the mind? The answer is negative. Memory traces can give rise to conscious and introspectable memory states, but they are removed from

consciousness. They are the opaque entities that explain the coming about of (potentially) transparent memory states. What we are able to become aware of, are not the traces, but the states of recall they give rise to. Memory traces are opaque intentional states that represent past events or experiences, and, when activated, can give rise to conscious thought and conscious behavior.[12]

3.3. Conceptual Evidence

Neither the third-person data provided by today's neurobiology, nor the first-person data of introspective reports, verify the trace hypothesis. In light of this fact, some philosophers have tried to establish the existence of traces by an a priori argument. They argue that the concept of a trace is implied by the notion of remembering. Among those who hold that the existence of memory traces is knowable a priori is Martha Kneale, who writes:

> It is involved in the ordinary notion of memory or recollection that the memory event should have as a part-cause the occurrence of the event recollected. This is what makes it so easy for us to accept the story of brain traces as a physiological condition of remembering. They fill in the gaps in the causal chain which is felt to be necessary to explain recollection (1972, p. 2).

It takes only a little thought to see that the concept of a trace is not *implied* by the concept of remembering. Our language of memory makes no reference to physical traces. Traces are not what we mean when we talk about remembering, any more than the secretion of digestive juices is (part of) what we mean by 'eating', even though this is something that takes place when we eat.[13] How else is it possible that children learn what 'remembering' and 'eating' mean long before they learn anything about biology? Traces are a feature of the dominant *hypothesis* about memory, rather than a feature of the *concept* of memory. What is implicit in the language of memory is only that there be some form of causal link between the past and the present representation. Just what sort of link it is has to be established by other means.

4. Mnemic Causation

After having realized that the empirical evidence for the existence of memory traces is "not quite conclusive," Russell went ahead and proposed an account of memory causation which manages without the stipulation of memory traces. 'Mnemic causation' is the name of this account of memory causation.

While the theory of contiguous memory causation maintains that traces are the proximate cause of states of recall, the theory of mnemic causation holds that a past representation is directly causally relevant over a temporal distance. Both theories of causation are divided over the question of whether cause and effect may be separated by a time gap. The contiguity theory regards causation at a temporal distance to be impossible and therefore postulates traces that are produced by the past representation and persist into the present. The theory of mnemic causation, on the other hand, is prepared to accept that cause and effect don't have to be contiguous. Given mnemic causation, a past experience is not only "part of a *chain* of causes leading to the present event" but is (together with some retrieval cue[14]) "the *proximate* cause" of the state of recalling.[15] 'Mnemic causation' amounts to direct causal action at a distance in time.

In Charles Broad's *The Mind and its Place in Nature*, an entire chapter is devoted to the notion of a memory trace. Broad characterizes the difference between the theory of contiguous memory causation and Russell's theory of mnemic causation as follows:

> On the trace theory, if you were to take a cross-section of the history of the experient's body and mind anywhere between the past experience and the stimulus you would find something, viz., the trace, which corresponds to and may be regarded as the representative of the past experience. On Mr. Russell's theory [...] these intermediate slices, though relevant and necessary, would contain nothing which corresponds to and represents the past experience. [...] Although there is continuity between the *total* cause and the effect [...], yet there is no continuity between the effect and *each* independently necessary factor in the cause. The original experience is not joined on to the memory either directly; or by [...] some special persistent which represents it (1925, pp. 458–9).

In addition, Broad illustrates the difference between both accounts of causation by means of two diagrams (1925, pp. 444–5). Dots stand for momentary events, circles for memory images, crosses for persistent memory traces, full arrows for causal relations, and dotted arrows for cognitive relations. Moreover, '*e*' stands for a past event, '*t*' for a trace, '*s*' for a stimulus (or prompt), '*i*' for a memory image, and '*m*' for the memory of event *e*. The trace hypothesis is represented by figure 2. Here the past event *e* brings about a persisting trace *t* which, at some point, is activated by a prompt *s*, and produces the memory image *i*, which represents *e*. Russell's theory of mnemic causation is represented by figure 3. Here the past event *e* and the present prompt *s* together directly produce the memory image *i*, which represents the past event *e*.[16]

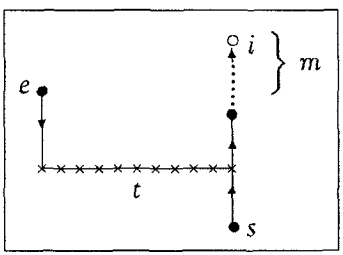

Figure 2

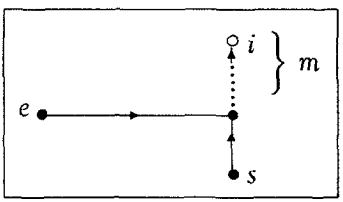

Figure 3

An undeniable advantage of Russell's theory of mnemic causation over the trace hypothesis is that it has the economy of postulating the existence of fewer relations of causal relevance among phenomena. Nevertheless, most critics find the notion of mnemic causation quite implausible. The principal objection to the idea of mnemic causation is that a cause cannot operate when it has ceased to exist, because what has ceased to exist is nothing. Without causal contiguity, a past event would somehow have to leap to the present to cause one's present memory activity. Broad formulates this objection as follows:

> According to the theory of mnemic causation my perception of a town which I visited last year literally produces a memory of this event whenever a suitable stimulus acts on me. But the perception is long past and is in no sense continued into the present. It has ceased to exist itself, and nothing now exists which can be regarded

> as a continuation of it. How then can it *do* anything now (1925, p. 452).

Notwithstanding its intuitive appeal, this objection to Russell's notion of mnemic causation is ultimately not convincing. There are two reasons that count against this objection.[17]

First of all, the difficulties surrounding the notion of causal action at a temporal distance rest on the *activity theory* of causation, i.e., the view that causes engage in an activity in order to bring about effects. Given the activity theory, a past experience cannot be the direct cause of a present state of recall, for the past experience cannot *do* anything, once it has ceased to exist. However, there are good reasons to reject the activity theory of causation. Causes do not act or operate. Russell declares, "[a] volition 'operates' when what it wills takes place; but nothing can operate except a volition" (1986, p. 183). Instead of the activity theory, Russell advocated the *uniformity theory* of causation. On this view, for *A* to cause *B*, all that is required is that whenever *A* is fulfilled, *B* happens, and whenever *B* happens, *A* has been fulfilled. Whether or not *A* and *B* are contiguous is irrelevant. Russell emphasizes that "*any* case of sufficiently frequent sequence will be causal" (1986, p. 185; 1995, pp. 88–9, 93). Once causation is defined in terms of regular sequences, nothing stops us from countenancing causal action at a distance in time.[18]

Secondly, the idea that a cause can bring about an effect after it has ceased to exist is not unique to the theory of mnemic causation. On the theory of contiguous causation, cause and effect are also separated by a finite time interval. The interval is much smaller than in the case of mnemic causation, but it must exist, for otherwise cause and effect become indistinguishable.[19] Thus, if Broad's objection were a good one, it would refute both the theory of mnemic causation and the theory of memory traces.

Before concluding the exposition of the notion of mnemic causation, I should mention that Ludwig Wittgenstein had a lot of sympathy for Russell's proposal. It is a common strategy of Wittgenstein to argue against certain seemingly natural assumptions that we make. A popular view of his time was Wolfgang Köhler's *principle of isomorphism*, i.e., the idea that psychological and physical processes are two

facets of a single process. The gestalt psychologist Köhler maintained "that the structural properties of experiences are at the same time the structural properties of their biological correlates" (1940, p. 109). What is of interest in the present context is not Wittgenstein's critique of the principle of isomorphism, but some general remarks regarding mental causation that are a by-product of his treatment of isomorphism. In *Zettel*, Wittgenstein writes:

> 610. I saw this man years ago: now I have seen him again, I recognize him, I remember his name. And why does there have to be a cause of this remembering in my nervous system? Why must something or other, whatever it may be, be stored up there *in any form?* Why *must* a trace have been left behind? Why should there not be a psychological regularity to which *no* physiological regularity corresponds? If this upsets our concept of causality then it is high time it was upset.
> 611. The prejudice in favor of psychological parallelism is a fruit of primitive interpretations of our concepts. For if one allows a causality between psychological phenomena which is not mediated physiologically, one thinks one is professing belief in a gaseous mental entity.
> 613. Why should there not be a natural law connecting a starting and a finishing state of a system, but not covering the intermediary state? (Only one must not think of *causal efficacy*).[20]

These passages are clearly directed against the need for a correlation between brain processes and thoughts, i.e., thoughts need not be reduced to brain processes. But this is not the only target. Wittgenstein dismisses not only the requirement for isomorphic correlation but also the requirement of mediative causality. That Wittgenstein was toying with the idea of direct causation at a temporal distance also becomes also apparent when he writes, "[t]here is something like action at a distance here—which shocks people. The idea would revolutionize science."[21]

5. The Explanatory Force of Memory Traces

In the previous section, we saw that Russell's notion of mnemic causation does not face any crucial difficulties. It represents a viable

alternative to the theory of memory traces. Moreover, in section 3, we saw that today's neurobiology yields no conclusive evidence for the existence of memory traces. Thus, the arguments for and against memory traces seem to balance one another. The notion of mnemic causation flies in the face of the everyday contention that the time span between cause and effect has to be much shorter than the interval between learning and retention. Yet mnemic causation has the economy of postulating fewer relations of causal relevance between a past experience and its subsequent recall than the trace theory. The trace hypothesis, on the other hand, conforms with our natural assumption that cause and effect must be temporally contiguous, or very nearly so. Yet so far there is no *decisive* empirical proof for the thesis that memory causation operates through engrams.

The apparent tie between the theory of mnemic causation and the theory of contiguous causation is resolved as soon as we treat memory traces as *theoretical entities*, i.e., devices introduced in the context of a theory to explain some more accessible phenomenon. The ontological status of the concept 'memory trace' is like that of 'equator' or 'centre of gravity'. When memory traces are taken to be theoretical constructs, to find fault with the theory of memory traces is, at least in part, to cast doubt on our need to postulate traces in order to account for remembering. Conversely, arguments in favor of the trace hypothesis have to demonstrate that the concept of a trace allows us to explain certain features of the intuitive notion of memory. The goal of this paper is to demonstrate that the stipulation of memory traces is indispensible for analyzing memory causation.

Assuming the existence of memory traces, the causal process underlying memory consists of *three* elements: encoding, storage, and retrieval. In encoding, experiences and thoughts bring about memory traces. In storage, the information is communicated from one trace to another. In retrieval, traces (together with cues) bring about states of recall. According to Russell's notion of mnemic causation, however, the causal process underlying memory consists of only *two* elements: storage and retrieval.

Russell omits information encoding (i.e., the formation of traces on the basis of experiences and thoughts) because he questions the existence of memory traces. Apart from the fact that the trace the-

ory *does*, while the theory of mnemic causation does *not*, assume that information is encoded in traces, both theories present different interpretations of memorial retention. According to the trace theory, information is stored in traces. In Russell's view, however, it is the original representation (or experience) itself that is somehow stored away until it is reactivated by the appropriate retrieval cues. Finally, the accounts of information retrieval differ. In Russell's view, cues activate the original experiences. In the trace hypothesis, cues activate traces that are derived from the original experiences.

My strategy is to show that the concept of a memory trace and the tripartite distinction of the memory process that goes with it allow us to explain certain features of the intuitive notion of memory which cannot be accounted for by the theory of mnemic causation.

6. Information Storage

Consider the following scenario (adopted from Martin and Deutscher 1966, pp. 180–1): at t_1, Oscar is involved in a car accident. At t_2, Oscar tells Bert about the accident. Then, at t_3, Oscar has a second accident as a result of which he forgets everything about the first accident. When, at t_4, Bert notices that Oscar can no longer remember the first accident, he tells him the details that Oscar had told him at t_2. Oscar hears Bert tell him about his first accident, and at t_5, Oscar retells this account of his accident. Although Oscar's retelling the story of his first accident at t_5 is causally related to his having witnessed the accident at t_1, intuitively he does not remember the accident. The reason his retelling of the first accident does not qualify as remembering is that the causal chain connecting his initial experience and his subsequent retelling follows an *external loop*. Now which of the two accounts of memory causation at hand—mnemic causation or the trace theory—is more apt to rule out cases of memorial retention resting on deviant causal chains?

The trace theory's explanation for why Oscar's retelling of his first accident does not qualify as remembering runs as follows: the intentional object of a memory report is determined by the proximate cause of the traces which bring about the state of recalling. The in-

tentional object of Oscar's recounting at t_5 is not his experience of the first accident at t_1, since the traces produced by this experience have been erased by the second accident. Instead, the traces that are responsible for Oscar's recounting of his first accident at t_5 are derived from the fact that at t_4 Bert told Oscar about this accident. Thus, rather than remembering the accident Oscar remembers what Bert told him about the accident.

How then would Russell deal with the example of memorial retention following an external loop? As was explained in section 4, the theory of mnemic causation demands that the original experience is the *proximal* cause of the state of recounting. Remembering is said to be *direct* causal action at a distance in time. According to Russell, Oscar does not remember the first accident because the causal chain connecting his experience of the first accident at t_1, and his retelling of this accident at t_5 is *indirect*. It is indirect for it encompasses Bert's memory.

We record that the trace theory and the theory of mnemic causation are equally suited to exclude memorial retention following an external loop. Nevertheless, I believe that the trace theory's account of retention is superior to that of the theory of mnemic causation. To see this we only need to raise the question of where and how information is supposed to be stored, if not in traces.

In the *Analysis of Mind*, Russell leaves open how we should conceive of information storage within the framework of mnemic causation. Clearly, without the assumption of memory traces, information storage cannot be conceived as a causal *process*. But how can we make sense of the idea of trace-free retention of information? To my mind, the only way of spelling out the idea of trace-free retention is in terms of dispositions. According to the dispositional account of retention, all that is required to remember an event is that, in virtue of having witnessed the event, one acquired a disposition to represent it, a disposition which one retained and now exercises by thinking of the event. Remembering that P would be simply one's persisting disposition to produce tokens of the thought that P in certain circumstances. There would be no need for a causal connection between one's past witnessing and one's present representation.[22]

The dispositional analysis of information storage appears coher-

ent only as long as one doesn't ask what is involved in the retention of dispositions such as the disposition of recounting a past event. When this question is raised, Russell would have to concede that the trace-free notion of memorial retention cannot provide an answer. But if the cognitive process underlying trace-free retention of information cannot be explained, memory becomes a magical process bearing some resemblance to telepathy and clairvoyance. For, as Max Deutscher declares, in claiming "the continuity of capacities [and dispositions], [we] are always committed to the continuity of *some* processes adequate to the continuity" (1989, p. 62). Thus, the account of information storage based on the trace theory is not only more convincing than the account of information storage based on mnemic causation but, when spelled out, the latter collapses into the former. The very idea of memorial storage calls for the stipulation of traces.

7. Information Encoding and Retrieval

Let's start with the science fiction movie *Total Recall*. In *Total Recall* people can travel to other planets without leaving home. The lead character, played by Arnold Schwarzenegger, cannot afford his dream vacation to the planet Mars. So he contacts the company *Rekall Incorporated* to have exotic memories of traveling to Mars implanted into his brain. *Rekall Incorporated* prides itself in creating low-cost vacation memories; as an added bonus, there is no chance that your luggage will be lost. Unfortunately for Arnold, things don't work out as planned. He discovers that he has already lived on Mars, where he worked in a corrupt government. As a result, part of his memory has been removed to keep the corruption secret. In the process, Arnold becomes thoroughly confused about what is real and what is not.

What is particularly intriguing about the movie *Total Recall* is the mere idea that we might one day possess the technical capabilities to artificially create memories in the mind of a person who would then experience those pseudo-memories as indistinguishable from genuine recollections of the past.[23] Instead of artificially created traces one

could also imagine that traces are taken from a person who has traveled to Mars and are implanted in Arnold's brain.[24]

It is beyond doubt, I reckon, that Arnold Schwarzenegger does not *remember* having lived on Mars—not even if he did, in fact, live on Mars. Surgically implanted memory traces cannot give rise to genuine memory, not even if these traces 'represent' propositions which are true, or if they 'represent' events one did experience in the past. The reason implanted traces do not bring about genuine memory states is that there is no causal connection between the content of the original experience and the content of the subsequent recall. To drive home this point, suppose that on his first visit to Mars, of which, initially, he has no recollections, Arnold thought, say, that Mars is pretty. After having received artificial memories from *Rekall Incorporated*, it might seem to Arnold as if he remembers having believed that Mars is pretty. This memory-impression does not qualify as remembering since his past belief that Mars is pretty is not the reason for his present belief that Mars is pretty. He doesn't believe that Mars is pretty now *because* he used to believe that Mars is pretty.

How do the theory of mnemic causation and the theory of memory traces, respectively, deal with the example at hand? Are both theories capable of capturing our intuition that Arnold does not remember having traveled to Mars? Given the theory of memory traces, it is easy to differentiate between naturally caused traces and artificially caused traces. Thus, it is easy to distinguish between genuine memories, on the one hand, and Arnold's spuriously caused pseudo-memories, on the other. Traces capable of giving rise to genuine states of remembering are derived from representations (or experiences) of the very subject bearing the traces.

But what about Russell's conception of mnemic causation? Is the notion of mnemic causation capable of telling real memories apart from Arnold's sham memories? On what basis can we deny Arnold memories of Mars once we question the existence of traces? I believe that there *is* a way for Russell to amend his theory of mnemic causation in order to rule out Arnold-type cases. All he needs to do is to add the claim that memory implies personal identity:[25] the person experiencing P at t_1 must be numerically identical with the person recalling P at t_2. Given this additional constraint on mnemic cau-

sation, Arnold does not qualify as remembering having traveled to Mars.

In sum, our first conclusion is negative: the stipulation of memory traces isn't necessary for distinguishing the causal process underlying remembering from pseudo-memories based on implanted traces. The theory of memory traces and Russell's notion of mnemic causation both can rule out such spurious causal routes.

7.1. Counterfactual Dependence

My argument to the effect that the postulation of traces is indispensable for analyzing memory causation has the form of a reductio: suppose we question the existence of traces and the tripartite distinction of memory causation that goes with it. Without the tripartite distinction, the causal process underlying remembering presents itself as an indivisible process that possesses the same strength anywhere between the original representation and the subsequent recall. There is no room to account for varying causal strengths between the different sections of the causal chain connecting the original representation and the recall. And this is where the problem lies. For the strength of the causal relation characteristic of information encoding differs from the strength of the causal relation characteristic of information retrieval. The causal relation of information encoding supports counterfactual conditionals while the causal relation of information retrieval does not. Since Russell's theory of mnemic causation is unable to accommodate for differences in causal strengths between encoding and retrieval, it is either too broad or too narrow. If mnemic causation is spelled out in terms of counterfactuals, it yields an interpretation of the retrieval process which is too narrow; and if it is not analyzed in terms of counterfactuals, it is too broad to provide a convincing interpretation of the encoding process. Given that it is impossible to explain remembering using only one notion of causal dependence, we are forced to introduce the tripartite distinction of encoding, storage and decoding. But this distinction, in turn, calls for the stipulation of memory traces. Hence, traces are an indispensable part of the analysis of memory.

The crucial premise of my argument for the explanatory force of the trace hypothesis consists in the twofold claim that the causal relation underlying information encoding supports counterfactuals, while the causal relation underlying information retrieval doesn't. Let's start with the former claim.

We all know that our mood can taint our (apparent) memory.[26] Suppose that due to a severe depression, Oscar is convinced that no one likes him. Whenever he has a pleasant encounter (which happens rarely enough), the experience gets transformed in his memory, so that, later on, it seems to him as if he had felt the meeting was unpleasant. Oscar is unaware of his memory's bias. Now suppose that at t_1 Oscar has an encounter that he feels is unpleasant. At t_2 he seems to remember that the meeting at t_1 was unpleasant. Thus, the memory claim is true and, what is more, there is a causal relation between his experience at t_1 and his memory claim at t_2. But does Oscar remember having had an unpleasant encounter at t_1? The answer, I take it, is negative. The reason Oscar doesn't remember that the encounter at t_1 was unpleasant is that if the encounter had been pleasant, he would still believe that it was unpleasant. It is pure luck that the mood he takes himself to remember corresponds to his original mood. This example indicates that causation alone isn't sufficient for information transmission from original experiences to their subsequent recalls; possible causal relations (or counterfactual dependencies) also contribute to determining information encoding.

Contrary to information encoding, the causal relation underlying information retrieval doesn't support counterfactuals. To see that the causal necessity of information retrieval is considerably weaker than that of information encoding, consider the following thought experiment (adopted from Martin and Deutscher 1966, p. 186): suppose Bert takes a potent hypnotic drug (such as sodium amytal) which causes him to become suggestible to all sorts of credible promptings—true and false—concerning what he has done and seen.[27] Apart from permanently inducing in him a suggestible state, the drug has no other unusual psychological impact; it doesn't cloud his consciousness nor does it affect his ability to remember, when not prompted. Thus, the presence of the suggestible state, by itself, doesn't rule out the possibility of remembering; it only does so when Bert is prompted.

When Bert remembers that *P* without being prompted, the following causal statement is satisfied:

> Bert's representing that *P* at t_1 causes his recounting that *P* at t_2.

But this causal statement may not be translated into the following counterfactual statement:

> Bert would not recount that *P* at t_2 unless he represented that *P* at t_1.

The reason the causal statement may not be rephrased in terms of a subjunctive conditional is that it is possible that Bert recounts that *P* because he was prompted. When prompted, his suggestible state makes Bert recount that *P*, regardless of whether he possesses the relevant memory traces.

Cases of enhanced suggestibility show that whether the causal process underlying information retrieval qualifies as remembering is determined by what *actually* causes the state of recounting, and not by how it would have been caused, if things had gone differently. If there had been independent sufficient causation held in reserve on a deviant route (e.g., a retrieval cue), this would not affect the question whether an instance of free recall (i.e., cue-independent recall) qualifies as memory.[28]

In sum, the crux of Russell's theory of mnemic causation is that it must assume that the causal strength is the same throughout the causal chain connecting the original representation and the recall. For this reason mnemic causation turns out to be either too narrow or too broad. If mnemic causation is spelled out in terms of counterfactuals it yields an interpretation of the retrieval process which is too narrow; if it is not analyzed in terms of counterfactuals it is too broad to provide a convincing interpretation of the encoding process. Therefore, the stipulation of memory traces is indispensable for analyzing memory causation.[29]

References

Anderson, J. R. 1995: *Cognitive Psychology and its Implications*. New York, W. H. Freeman.

Annis, D. B. 1980: Memory and Justification. *Philosophy and Phenomenological Research* **40**: 324–33.

Ayer, A. J. 1972: *Probability and Evidence*. London, Macmillan.

Baddeley, A. 1990: *Human Memory. Theory and Practice*. Boston, Allyn and Bacon.

Beauchamp, T. L., Rosenberg, A. 1981: *Hume and the Problem of Causation*. New York, Oxford University Press.

Bernecker, S. 2001: Impliziert Erinnerung Wissen?. In T. Grundmann (ed.), *Erkenntnistheorie. Positionen zwischen Tradition und Gegenwart*, Paderborn, Mentis, 145–64.

Bower, G. H. 1981: Mood and Memory. *American Psychologist* **36**: 129–48.

Broad, C. D. 1925: *The Mind and its Place in Nature*. London, Kegan Paul.

Deutscher, M. 1989: Remembering 'Remembering'. In J. Heil (ed.), *Cause, Mind, and Reality. Essays Honoring C. B. Martin*, Dordrecht, Reidel, 53–72.

Ducasse, C. J. 1951: *Nature, Mind, and Death*. La Salle, Open Court.

Dudai, Y. 1989: *The Neurobiology of Memory*. Oxford, Oxford U. P.

Flage, D. E. 1985: Hume on Memory and Causation. *Hume Studies* **10**, Suppl., 168–88.

Ginet, C. 1975: *Knowledge, Perception, and Memory*. Dordrecht, Reidel.

——. 1988: Memory Knowledge. In G. H. R. Parkinson (ed.), *Handbook to Western Philosophy*, New York, Macmillan, 159–78.

Gloor, P., Olivier, A., Quesney, L.F., Andermann, F., Horowith, S., 1982: The Role of the Limbic System in Experimental Phenomena of Temporal Lobe Epilepsy. *Annals of Neurology* **12**: 129–44.

Heil, J. 1978: Traces of Things Past. *Philosophy of Science* **45**: 60–72.

Hume, D. 1975: *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Ed. by L. A. Selby-Bigge, 3rd edition by P. H. Nidditch, Oxford, Clarendon Press.

——. 1978: *A Treatise of Human Nature*. Ed. by L. A. Selby-Bigge, 2nd edition by P. H. Nidditch, Oxford, Clarendon Press.

Kneale, M. 1972: Our Knowledge of the Past and Future. *Proceedings of the Aristotelian Society* **72**: 1–12.

Köhler, W. 1940: *Dynamics of Psychology*. New York: Liveright.

Locke, J. 1975: *An Essay Concerning Human Understanding*. Ed. by P. H. Nidditch, Oxford: Clarendon Press.

Malcolm, N. 1963: *Knowledge and Certainty*. Ithaca, Cornell U. P.
——. 1977: *Memory and Mind*. Ithaca, Cornell University Press.
Martin, C. B., Deutscher, M. 1966: Remembering. *Philosophical Review* **75**: 161–96.
Mayes, A. R. 2001: Aware and Unaware Memory. Does Unaware Memory Underlie Aware Memory? In C. Hoerl and T. McCormack (eds.), *Time and Memory. Issues in Philosophy and Psychology*, Oxford, Clarendon Press, 187–211.
Nagel, E. 1961: *The Structure of Science*. New York: Harcourt, Brace and World.
Parfit, D. 1984: *Reasons and Persons*. Oxford, Clarendon Press.
Penfield, W. 1958: Some Mechanisms of Consciousness Discovered During Electrical Stimulation of the Brain. *Proceedings of the National Academy of Sciences* **44**; 51–66.
Reason, J. T., Mycielska, K. 1982: *Absent Minded? The Psychology of Mental Lapses and Everyday Errors*. Englewood Cliffs, Prentice Hall.
Reid, T. 1868: *The Works of Thomas Reid*. Ed. by W. Hamilton, 2 volumes, Edinburgh, MacLachlan and Stewart.
Rose, S. 1992: *The Making of Memory*. London, Bantam Press.
Rumelhart, D. E., Norman, D. A. 1981: A Comparison of Models. In G. E. Hinton and J. A. Anderson (eds.), *Parallel Models of Associative Memory*, Hillsdale, Erlbaum, 1–7.
Russell, B. 1959: *The Problems of Philosophy* (1912). Oxford, Oxford University Press.
——. 1986: On the Notion of Cause (1912). In his *Mysticism and Logic*, London, Unwin, 172–99.
——. 1995: *The Analysis of Mind* (1921). Introduction by T. Baldwin, London, Routledge.
Semon, R. 1909: *Die mnemischen Empfindungen in ihren Beziehungen zu den Originalempfindungen*. Leipzig, Engelmann.
Shoemaker, S. 1970: Persons and their Pasts. *American Philosophical Quarterly* **7**: 269–85.
Shope, R. 1973: Remembering, Knowledge, and Memory Traces. *Philosophy and Phenomenological Research* **33**: 303–22.
Siebel, M. 2000: *Erinnerung, Wahrnehmung, Wissen*. Paderborn, Mentis.
Smart, J. J. C. 1991: Sensations and Brain Processes. In D. M. Rosenthal (ed.), *The Nature of Mind*, New York, Oxford University Press, 169–76.
Squire, L. R. 1987: *Memory and Brain*. New York, Oxford University Press.
Squires, R. 1969: Memory Unchained. *Philosophical Review* **78**: 178–96.

Stein, D. G., Glasier, M. M. 1995: Some Practical and Theoretical Issues Concerning Fetal Brain Tissue Grafts as Therepy for Brain Dysfunctions. *Behavioral and Brain Sciences* **18**: 36–45.

Stroud, B. 1977: *Hume*. London, Routledge.

Wiggins, D. 1980: *Sameness and Substance*. Cambridge, Cambridge University Press.

Williams, B. 1973: *Problems of the Self. Philosophical Papers 1956–72*. Cambridge, Cambridge University Press.

Wittgenstein, L. 1967: *Zettel*. Ed. by G. E. M. Anscombe and G. H. von Wright, Berkeley, University of California Press.

——. 1980: *Remarks on the Philosophy of Psychology*. Ed. by G. E. M. Anscombe, 2 volumes, Chicago, University of Chicago Press.

——. 1993: *Philosophical Occasions: 1912–51*. Ed. by J. C. Klagge and A. Nordmann, Indianapolis, Hackett.

Zemach, E. M. 1983: Memory: What It Is, And What It Cannot Possibly Be. *Philosophy and Phenomenological Research* **44**: 31–44.

Keywords

causation; connectionism; contiguity; dispositional belief; memory; memory trace; Russell; simultaneity paradox; subdoxastic state; Wittgenstein

Sven Bernecker
Institut für Philosophie
Universität München
Geschwister-Scholl-Platz 1
80539 München, Germany
Bernecker@lrz.uni-muenchen.de

Notes

[1] In (2001) I have argued that remembering that *P* implies neither believing nor knowing. Belief and knowledge supervenes some but not all cases of remembering. What passes into memory may be nothing but a subconscious thought or a fleeting experience. For this reason I call the input of the memory process *representation* or *experience* rather than a 'thought', 'belief', or 'knowledge' (cf. footnotes 9 and 10).

[2] 1978, p. 75, cf. pp. 170, 173. A number of Hume scholars have argued that Hume did not take contiguity to be a necessary condition for causation

(cf. Flage 1985, pp. 179–86; Stroud 1977, pp. 43–4). Hume admitted that we do not get an impression of contiguity every time we observe a pair of objects that we take to be related as cause and effect. In such case, we only 'presume' that there is contiguity nevertheless. In claiming that we may merely 'suppose' contiguity to be essential to causation, "according to the general opinion", Hume can be read to be giving little more than an enumeration of the common assumptions regarding causation. Nothing he said demonstrates that 'A causes B' implies 'A and B are contiguous'. Furthermore, in the *Enquiry Concerning Human Understanding* no appeal is made to contiguity in the definition of causation. Beauchamp and Rosenberg (1981, pp. 194–5), however, maintain that Hume used 'succession' and 'contiguity' as near synonyms. They argue that the fact that the *Enquiry* only incorporates succession into the definition of 'cause' and makes no mention of contiguity does not mean that Hume believed that causation may not be contiguous.

[3] "On the Notion of Cause" was Russell's presidential address to the Aristotelian Society in 1912. In his Lowell Lectures delivered in Boston in 1914 he held essentially the same views.

[4] Given probabilistic causation, one event's causing another does not require that the former determines the latter, but only that it makes it more probable than it would otherwise have been. For smoking to cause lung cancer it doesn't have to be the case that all smokers get cancer, only that the conditional probability of cancer, given smoking, is greater than the probability of lung cancer in general. That is, an event A is a probabilistic cause of an event B if the probability of the occurrence of B, given that A has occurred, is greater than the antecedent probability of B: $P(B/A) > P(B)$. This implies that event B must be more probable given the presence rather than the absence of the cause: $P(B/A) > P(B/\neg A)$.

[5] I owe this point to Peter Baumann.

[6] Cf. Dudai (1989) and Squire (1987). Shoemaker (1970, p. 282), Wiggins (1980, pp. 219–20), and Williams (1973, pp. 75–6) argue that the causal theory of memory conflicts with the concept of a mind as an immaterial substance. According to the causal theory, for someone to remember something there must be some appropriately characterized causal chain which links her present representation to her past observation. The idea of this causal chain is unpacked into the idea of a physical trace. The crucial step of the argument is the claim that the idea of a structurally complex memory trace belonging to an immaterial substance is incoherent. Memory traces have to have a physical basis. The rest of the argument is a simple reductio. Let's suppose dualism were true and we existed independent

of our bodies. We could not have any memories of experiences dating from our embodied state, for there is nothing to play the role of the causal chain required by the causal theory of memory. A disembodied person could not be said to remember at all; and given that the ability to remember is an important aspect of what it means to be a person, it follows that it is impossible for a person to be disembodied. Hence, given the causal theory of memory, the suggestion that a person existed independent of her body is incoherent.

[7] I use 'fact memory' as synonymous with 'propositional memory'. Some authors use the term 'fact memory' (or 'factual memory') as the counterpart to 'ostensible memory'. According to this usage, fact memories imply truth while ostensible memories do not. All four kinds of memories that I differentiate—object-, property-, event-, and fact memory—imply truth. In the case of fact memory this is obvious. Object memories imply truth because "I remember the dog Fido" implies that there is something (Fido) that I remember. The same goes for the property- and event memory.

[8] Attributions of fact memory yield referentially opaque contexts, while attributions of object-, property-, and event memory are referentially transparent. In the statement "S remembers Marilyn Monroe" you can replace 'Marilyn Monroe' by 'Norma Jean Mortenson' without changing the truth value of the statement. However, from "S remembers that Marilyn Monroe is blond" and "Marilyn Monroe is Norma Jean Mortenson" it doesn't follow that "S remembers that Norma Jean Mortenson is blond".

[9] Other labels for subdoxastic states are 'implicit knowledge', 'tacit knowledge', 'proto-knowledge', 'unconscious knowledge', and 'subpersonal states'. The problem with characterizing subdoxastic states as 'unconscious' is that unconscious (or subconscious) states can, in principle, be made conscious while subdoxastic states cannot. The problem with the label 'knowledge' for subdoxastic states is that the information transmitted by them may be both false and unwarranted. Knowledge, however, implies truth and justification.

[10] Since remembering doesn't imply knowing, Siebel concludes that subdoxastic states may transmit false information (2001, pp. 236–57). I disagree. Of course, subdoxastic states may propagate 'misinformation', yet the states of recall based on these subdoxastic states cannot classify as memories, for memory implies truth. It is not the truth-condition but the justification-condition and the belief-condition which distinguishes remembering from knowing.

[11] 1995, p. 85. "I do not wish to urge that [mnemic] causation is ultimate, but that, in the present state of our knowledge, it affords a simplification,

and enables us to state laws of behaviour in less hypothetical terms than we should otherwise have to employ" (ibid).

[12] An immediate consequence of the fact that traces are removed from consciousness is that I cannot tell, by reflection alone, whether the state I am in is a state of remembering rather than of perceiving or imagining. For to know this I would have to be able to rule out the possibility that the state I occupy wasn't caused by a memory trace. But this I cannot do by reflection if traces are in principle unconscious.

[13] This analogy is borrowed from Malcolm (1963, p. 237). The critique of the idea that memory traces are knowable a priori echo some early objections to the psychophysical identity theory. When the identity theory was first advanced, it was objected that if mental states are identical with brain states how had this fact eluded attention for so long (cf. Smart 1991, p. 171). Identity theorists replied by pointing out that psychophysical identities are something we discover from observation and experience, not something that could be ascertained a priori or by merely investigating the meaning or concepts involved. The concept of, say, pain and the concept of C-fiber excitations are distinct and independent concepts, and this explains how it is possible for someone to know a lot about pains but nothing about C-fiber excitations. Psychophysical identities are empirical truths (assuming they are truths at all) which depend on scientific research. For this reason, the status of psychoneural identities is like that of theoretical identities in the sciences, e.g., 'temperature in gases is mean molecular kinetic energy', 'lightning is electrical discharge' and 'water is H_2O'.

[14] Retrieval cues are snippets of information that allow us to access traces. To be prompted means to be re-exposed to some relevant information. Prompting can occur intentionally or unintentionally and it can occur with or without our awareness. For example, a person is prompted if she just happens to read a fictional story that by chance matches something in her own past, or if she sees some event very much like another that she previously saw, or sees an object in much the same state as she saw it previously. Smells and tastes are particularly powerful cues. Probably the most famous literary illustration of this fact comes from Marcel Proust's *Remembrance of Things Past* where he describes how the taste and smell of a madeleine cake soaked in lime tea brings back with enormous vividness memories of his childhood. Apart from smells and tastes, the two most common kinds of cues are verbal and visual reminders. Verbal reminders can consist of just one word or of an entire narrative. Analogously, visual reminders can be single pictures or a series of pictures. Verbal reminders are often richer than visual reminders that usually touch the original experience at only one point.

[15] Russell 1995, pp. 78, 85. Russell gives the following illustration of mnemic causation: "you smell peat-smoke, and you recall some occasion when you smelt it before. The cause of your recollection, so far as hitherto observable phenomena are concerned, consists both of the peat-smoke (present stimulus) and of the former occasion (past experience). The same stimulus will not produce the same recollection in another man who did not share your former experience, although the former experience left no *observable* traces in the structure of the brain. According to the maxim 'same cause, same effect', we cannot therefore regard the peat-smoke alone as the cause of your recollection, since it does not have the same effect in other cases. The cause of your recollection must be both the peat-smoke and the past occurrence" (1995, pp. 78–9). Thus it is the past olfactory experience and not some present representative of the past experience (like a trace) that causes the recollection. The past experience is the proximal cause of the recall.

[16] Both diagrams presuppose that images are a necessary component of remembering. In the *Analysis of Mind* Russell holds that the intentional object of memory is not the past event itself but a present image that reproduces some past sense experience. Once memory is defined in terms of image-reproductions, we need some way of distinguishing memory images from pure imagination. Russell tries to solve this problem by suggesting that memory images are distinguished from other images by two feelings that accompany them: 'feelings of familiarity' that lead us to trust the images, and 'feelings of pastness' that lead us to refer them to some time in the past (1995, p. 163). In the *Problems of Philosophy* (pp. 114–5), which appeared nine years before the *Analysis of Mind*, Russell maintained that our awareness of the past is direct. A memory-impression consists in the rememberer's directly experiencing the past event itself, without a representative intermediary such as an image. A few years after the publication of the *Problems*, Russell gave up direct realism and advocated indirect realism.

[17] This isn't the only objection launched against the notion of mnemic causation. Broad argues that mnemic causation doesn't give rise to "an ultimate causal law" but, at best, to "an empirical generalization of the very crudest kind" (1925, p. 459). The reason is that we expect a causal law to specify the interval between the cause and the effect. The theory of mnemic causation, however, cannot specify any constant time-relation between a past experience and the subsequent recall but instead has to allow for temporal gaps of various sizes. Given the nomological character of causality, Broad concludes, Russell's notion of mnemic causation is causation only by name. I am not convinced by Broad's objection, for I believe that there might be

causal laws that do not specify the time interval between cause and effect. Consider an example by Shope (1973, p. 320): we can say that bending a copper wire more than ten times in less than fifteen seconds causes it to break without having to say how long the interval between the individual bendings has to be.

[18] A standard argument against the uniformity view of causation is the so-called *accidental generalization problem.* Constant conjunctions include accidental generalizations as well as lawlike regularities, whereas only the latter give rise to causal connections. Reid, for example, noted that day is invariably followed by night and night by day and yet neither is the cause of the other (1868, II, p. 627). Similarily, Ducasse (1951) noted that in infants the growth of teeth invariably follows, but is not caused by, the growth of hair. Innumerable constant conjunctions like these just happen to hold, and so do not constitute causal connections between events. This presents a serious problem for the regularity view, because no one has yet succeeded in distinguishing laws from accidental generalizations except in terms of natural necessity.

[19] See the discussion of the simultaneity paradox in section 1.1.

[20] These passages have also been published as § 905, 906, and 909 of the first volume of the *Remarks on Philosophy of Psychology.*

[21] 1993, § 411. In addition to Wittgenstein himself some of his disciples adopted the notion of mnemic causation: Annis (1980, p. 331), Ginet (1975, pp. 166–9; 1988, pp. 166–7), Heil (1978, pp. 68–9), Malcolm (1977, p. 187), and Shope (1973, pp. 317–22).

[22] Russell himself suggested a dispositional account of information storage: "[m]emories, as mental facts, arise from time to time, but do not, so far as we can see, exist in any shape while they are 'latent'. In fact, when we say that they are 'latent', we mean merely that they will exist under certain circumstances" (1995, p. 86). The most detailed defense of the kind of dispositional account of retention that I attribute to Russell has been given by Squires (1969).

[23] Whether mind transplants will ever become a medical possibility is uncertain. Brain transplants, however, are not only possible but real. Heads of monkeys have successfully been transplanted from one to another and brain tissue implantation is being explored in patients suffering degenerative disorders such as Parkinson's disease, Huntington's chorea, and Alzheimer's disease. Research on brain transplantation in humans is still very much in its infancy, but some experiments suggest that Parkinson patients who receive fetal brain tissue grafts improve (cf. Stein and Glasier 1995). But even if, some day, fetal brain tissue grafts should become a standard treatment for

degenerative disorders, this, by itself, wouldn't prove the possibility of mind transplants. For the fetal brain tissue is free of memory traces and by implanting it into an adult's brain no memories are being transmitted. The implanted tissue only serves the function of supplying additional storage space.

[24] Thought experiments concerning the duplication and the transplantation of minds, souls, and memories have a long history in philosophy. Locke imagined the soul of a prince slipping into the body of a cobbler (1975, p. 340). Shoemaker considered a person, called Brownson, who has the brain of Brown and the body of Robinson (1970, p. 282). Williams (1973, Ch. 1) and Parfit (1984, p. 220) invented the case of a brain whose information is duplicated and copied into another brain. And Zemach imagined a king who acquires the memories of his subjects by eating their brains (1983, p. 36). Thought experiments regarding mind transplants are taken to support the psychological continuity criterion of personal identity.

[25] If remembering presupposes personal identity, then memory cannot occur as an ingredient in a definition of personal identity. Yet the psychological continuity theory defines personal identity on the basis of memories. The (alleged) inconsistency goes under the label of the *circularity objection* to the psychological continuity criterion of personal identity. The customary reply to the circularity objection, originally given by Shoemaker (1970), is that while a definition of personal identity in terms of memory might be circular, one can define a more general concept of quasi-memory, which is not true but which is in all other essential respects identical with our ordinary concept of memory. Quasi-memory, like memory, is capable of yielding knowledge of the past that is based neither on empirical evidence nor testimony. Psychological continuity can then be redefined in terms of quasi-memory and the psychological continuity criterion of personal identity can be cleared of the accusation of circularity.

[26] Most psychologists agree that the mood can have an effect on memory. Psychologists differentiate between two types of effects a subject's mood might have on her ability to remember. First of all, there is *mood state dependency*, whereby anything experienced in a given mood will tend to be recalled more easily when that mood is reinstated, regardless of whether the material experienced in the mood is pleasant, unpleasant or neutral. Secondly, there is *mood-congruency*, whereby a given mood will tend to evoke memories that are consistent with that mood, hence, when sad we tend to recall sad events, even though encountered these during a period of happiness. By and large, the experimental evidence for mood-congruency is stronger than for mood-state dependency. See Bower (1981).

[27] Hypnosis and hypnotic drugs create a retrieval environment in which people are more willing than usual to call a mental experience a 'memory', and in which they express a great deal of confidence in both true and false memories. In extreme cases, individuals may become what has been called 'honest liars', believing strongly in implanted or imagined 'recollection'.

[28] Saying that the causal relation between memory traces and states of recall doesn't support counterfactuals is only a negative characterization. What would be a posititive characterization of the causal relation underlying information retrieval? Without being able to argue for this claim here, I suggest the following minimal condition for a state of recounting to qualify as remembering: the content of the relevant memory trace must be a necessary condition of some sufficient condition for the content of the state of recounting.

[29] An earlier version of this paper was presented at the Second Principia International Symposium in Florianópolis, Brazil in August 2001. I owe thanks to the audience at the symposium. Special thanks to Thomas Baldwin, Peter Baumann, Dorothea Debus, and Gary Hatfield.