16
Attacks on traces

A basic problem in understanding the organization of memory in a biological system is to understand how a vast quantity of information can be stored and recalled by a system composed of vulnerable and relatively unreliable elements, with no knowledge of where the information has been filed. (Leon Cooper 1980: 7)

16.1 A taxonomy of criticisms of trace theory

In chapter 15, seeking positive accounts of memory which dispensed with traces, I ran through a number of direct realist and Gibsonian objections to trace theory. Now I attempt a more extensive account of the criticisms which many readers will have felt rumbling beneath my descriptions of historical theories. Here, rather than hypothesising on critics' motives, I finally try to reply to their arguments.

Arguments against traces are bewilderingly abundant: many who hold quite different positive views about memory share broadly the same negative complaints. So, in moving towards an exhaustive taxonomy, I synthesise the polemics of Gibsonian ecological realists, phenomenological philosophers, Wittgensteinians, historians of philosophy, and direct-realist analytic philosophers under four broad headings: here I omit only more technical differences over reduction and levels of explanation, and a set of more extreme Wittgensteinian complaints, both of which I address elsewhere (Sutton 1993, 1995). The underlabouring character of this task is pronounced: only by a slow enumeration of possible reasons to reject traces and a careful explanation of the strategies of evasion by which distributed models avoid refutation can tedious philosophical overgrowth be cleared, and shared attempts to mix keen phenomenology with explanation be rejoined.

This survey of objections to traces catches the most troubling issues. But a range of less clearly articulated complaints lurks under them, around the discomfort critics feel about apparent scientism. Many critics cited in this chapter oppose the materialism and determinism of causal theories.¹ Many reject the

¹ Here I just ignore complaints about scientists' alleged reduction of all the diversity of remembering to one causal model, and about its unanalysable nature (Malcolm 1977; Bursen 1978). As I have shown throughout, science and complexity must coexist in memory theory, with social and personal levels of analysis as important as biological. Two suggestive treatments of the variety of memory phenomena which do not polemically rule out causal analyses are Deutscher 1989: 53–7 and Fentress and Wickham 1992: 1–40.
use of concepts unavailable to the subject in explanations of memory (Sanders 1985: 511–13). Some who do not explicitly address problems of reduction and levels mistake their preference for demonstration. Here I just give one example, as warning, of how difficult issues can simply be ignored.

Don Locke's introduction to the philosophy of memory begins:

Our first questions are obvious. What precisely is memory? What is it to remember something? These philosophical questions need to be distinguished from the similar questions that might be asked by a psychologist or neurologist. There are problems about how memory operates, about what goes on in our body or brain when we remember, about what physical and psychological factors help or hinder our remembering, and so on. These are the concern of the scientist, who investigates the functioning of the human capacity we call memory. But our question is different: we are asking what memory is in itself, what that human capacity consists in, however it may operate. We want to explain not the mechanisms of memory, but its nature; not how it works, but what it is. The question is, in effect, what it is we are talking about when we talk about memory. (Locke 1971: 1)

This is a defensible philosophical position: but to dissent from it is certainly not to stop doing philosophy. Locke makes, without argument, a specific claim about the autonomy of the mental from the levels which realise or constitute it. It cannot be ruled conceptually impossible that the nature of memory, what it is we are talking about, cannot be understood independently of its mechanisms, what it is independently of how it works.²

16.2 The role of empirical evidence

The importance of empirical evidence for or against the memory trace is far from clear. It did not, I suggested in part II, exclusively determine the fate of the animal spirits version of trace theory. Could the discovery of memory traces ever settle the issue? What would such a discovery involve? What is the status of the claim that traces exist? Is a trace theorist claiming that there must be a trace

² O'Connor and Carr (1982: 116) agree that 'questions about physical mechanisms are not problems for the philosopher' and that the 'scientific theories [of memory] established by experiment . . . are not part of philosophy' and do not 'have to be considered by philosophers': but they at least acknowledge that 'anything that philosophers may say about memory must at least be consistent with well established scientific theories'. I have suggested throughout this book that fruitful interactions as well as tensions have long characterised theories of memory. Johnson (1983: 377–8) argues that scientific evidence for the reconstructive rather than duplicatory nature of human (as opposed, say, to ant or beetle) memory is 'the most important problem' for the traditional analysis of memory as retained knowledge. Even those with untarnished faith in the analytic project need to heed Johnson's warning that 'no philosopher of memory can afford to ignore such scientific results, because they tell him what the phenomenon is, which he is trying to explain in philosophical terms'.

spanning the temporal gap between experience and remembering, and assuming that empirical research just fleshes out the particular way it is realised in the brain?

Critics argue on conceptual or 'logical' grounds against this whole empirical research programme. Bursen's Dismantling the Memory Machine is meant to demonstrate philosophically the absurdity of neuropsychology (1978: xii; cf. Krell 1990: xi). Maze describes as 'gratifying' the work on perception by the Gibsonians Turvey and Shaw, who show that 'empirical research cannot establish the truth of an indirect [representationist] theory of perception; of course, it cannot establish the truth of a direct one, either, since the issue is one of logical rather than empirical tenability' (1983: 78).

This is not the defensive position of the practitioner of conceptual analysis cordonning off areas of honest philosophical concern from the aggressive march of the sciences. It is, rather, polemical invasion, on a priori grounds, of the terrain of empirical science with arguments to show that alleged philosophical preconceptions of these sciences are ill-founded. For Sanders (1985: 507), 'the technical or theoretical accounts of memory that have been given are spurious': Wilcox and Katz (1981a: 238) claim that Gibsonian realism explains the phenomena of remembering, but that 'it does not, unlike the representative theory, accomplish [this] at the expense of logic'. If concepts like structural isomorphism, encoding, storage, and retrieval are logically incoherent, no amount of empirical research can save trace theory.

Some defenders of the trace seek, in reaction, to drain it of empirical content. The article on animal spirits in Chambers' Cyclopaedia noted that 'the existence of the animal spirits is controverted [disputed]: but the infinite use they are of in the animal oeconomy and the exceedingly lame account we should have of any of the animal functions without them, will keep the greatest part of the world on their side' (in Myer 1984: 103). Doubt about traces is to be dispelled by reference to the sheer difficulty of psychological explanation without them. Some modern trace theorists, seeking a notion of the trace which does not commit them to predictions about future neuroscience, put the existence of traces beyond the empirical. Johnson uses 'trace' 'in a content-less, functional way to mean simply 'that by means of which someone's past influences his present, in the course of remembering' (1983: 381, n. 6). And Deborah Rosen (1975: 3) develops a 'logical notion of the memory trace' which is 'independent of its physical realization', and is distanced 'from specialized scientific notions for which the logical notion provides only a philosophical underpinning' (compare Gomulicki 1953:vii).

Local and distributed traces, then, would be subspecies of this general notion which do involve specific claims about physical realisation. The bare, content-less concept of trace does some work: but unfortunately critics are tempted too easily to think that it still implies the crude sacks-in-storehouse
image. So arguments which purport to reject traces in general often in fact make a case only against localist traces.

Critics use counterexamples to remove the aura of necessity from the concept: ‘If I say, rightly, “I remember it”, the most different things can happen, and even merely this: that I say it’ (Wittgenstein 1974: para. 131; Malcolm 1977: chs. 6–10). Nothing here tells against distributed-memory theorists, who make no claims about what must happen in every episode of remembering. But it is natural to explain such a theory in terms which render the characterisation of the trace more specific. Critics fulminate against mechanism by repeating that the trace cannot be the memory (since, for example, it would be possible for someone who cannot remember an event nevertheless to have a trace of it).

An event leaves a trace in the memory: one sometimes imagines this as if it consisted in the event’s having left a trace, an impression, a consequence, in the nervous system. As if one could say: even the nerves have a memory. But then when someone remembered an event, he would have to infer it from this impression, this trace. Whatever the event does leave behind, it isn’t the memory. (Wittgenstein 1980: para. 220 3

The trace is not the occurrent remembering: but the trace theorist, if required, can agree to restrict the use of the word ‘memory’ to the act of remembering, or even, if philosophers decide that ordinary language so leads us, to restrict it further, as a ‘success-word’, to the act of remembering truly, although both historians and psychologists have provided ample reason to work on veridical and non-veridical ‘memory’ together. 4 But there can still be underlying dispositions which underpin the occurrent act of remembering.

So ‘the engram (the stored fragments of an episode) and the memory . . . are not the same thing’ (Schacter 1996: 70). Rather, the trace provides one kind of continuity between past event and present remembering. As Max Deutscher argues, it is not the stored trace which triggers or prompts present recall, nor is it the past event itself: the event recalled

3 Wittgenstein’s followers outdo him in attributing ludicrous views to their opponents. But in context, this remark (which can be compared to Gibson’s at the start of section 15.2 above) is aimed at the notion of memory traces as faithful reproductions or copies of past experience. Wittgenstein continues, mockingly: ‘The organism compared with a dictaphone spool; the impression, the trace, is the alteration in the spool that the voice leaves behind. Can one say that the dictaphone (or the spool) is remembering what was spoken all over again, when it reproduces what it took?’ There is a genuine target here, but it is Wittgenstein’s earlier self (Sutton 1993). Distributed models, in contrast, are not high fidelity reproduction systems.

4 Compare McCauley 1988: 129, complaining that the ‘epistemic honor’ we associate with ‘remember’ as opposed to ‘believe’ is irrelevant to arguments about the extent of reconstruction in memory. Likewise, facts about the justification of memory claims and ‘the concept of memory’, or that ‘our everyday verifications of whether some person does or does not remember that p are not bound up with any questions about what is and has been going on in this brain’ (Malcolm 1963: 237; also Munsat 1967: 9), do not mean that there are, in fact, no memory traces (Rosen 1975: 2).
did its work in the past in laying down a trace which now makes [it] possible
that the present mnemonic, the prompting, elicits recall. The ‘causation'
which is involved in memory is that of laying something down; it is a ‘causa-
tion' as sedimentation which makes a capacity possible later on. (Deutscher
1989: 61)\(^5\)

This formulation is sufficiently neutral that those opposed to literal copying
by traces of the structure of experience can find alternative senses for the
concept. ‘Causation as sedimentation’ does not imply that what is left after
experience must remain permanently fixed. Bartlett, theorising on the basis of
his evidence for the reconstructive and context-based nature of remembering,
wrote (1932: 211–12) that

though we may still talk of traces, there is no reason in the world for regarding
these as made complete at one moment, stored up somewhere, and then re-
excited at some much later moment. The traces that our evidence allows us to
speak of are interest-determined, interest-carried traces. They live with our
interests and with them they change.\(^6\)

Experiment, explanation, and theory continue even if context, culture, and
change are relevant to the memory trace. But for the trace to be a legitimate con-
struct, it must be connected to what it is a trace of only contingently. It must be a
‘distinct existence’ in order to stand in causal relations with the past, and it
must be identifiable independently of its nature as a trace of that past event
substrate of remembering, whatever that should be, cannot be the whole story.
Some independent account of what a trace is will ultimately be required of a
developed science. The philosophical and scientific projects cannot be iso-
lated, and at some point more specific claims about physical realisation are
required.

In chapter 4 I discussed the problem of action at a temporal distance, sug-
gest that animal spirits explanations of physical continuity in remembering
appealed to seventeenth-century natural philosophers who denied occult action
at a temporal distance. Their memory theory, in this respect, had to cohere with
the rest of their sciences: and the situation is similar now.\(^7\) But the postulation

---

\(^5\) Shope (1973: 321), complaining that arguments for traces should be scientific not
philosophical, acknowledges that the trace will not be the only causally relevant factor
involved in acts of remembering. The contexts of retrieval which interact with the trace, as
I have stated before, do not fall beyond the ambit of scientific psychology.

\(^6\) Causal theories of memory do not, as Zemach (1983: 32–3) complains, ‘dictate to science
what to discover in the brain’. Interesting scientific possibilities, like the ‘holographic
principles’ which Zemach cites as challenges to localised codes, are not rejecting memory
traces tout court, merely elucidating their (distributed) nature.

\(^7\) Malcolm, trying to expose the ‘confusion’ behind this route to trace theory, acknowledges
that ‘our strong desire for a mechanism (either physical or mental) of memory arises
from an abhorrence of the notion of action at a distance-in-time’ (1963: 238; compare
1977: ch. 6).
of memory traces in some neutral sense might be empirical without being empirical within psychology. The relevant empirical domain lies elsewhere, in the physical assumption of no action at a distance, of continuity between a cause and its effect when the effect is temporally subsequent. The assumption, verified in ordinary experience, is that mechanisms underlie apparent cases of direct action between temporally remote events (see Rosen 1975 on the ‘principle of temporal contiguity’). It may be mistaken: it is shared by folk physics and classical physics, but challenged in quantum physics. Challenges to it must meet certain constraints. Firstly, they require more than negative complaints about science’s lack of success in identifying the trace. Secondly, they require more than the bare point that the postulation of traces is not logically necessary.

Action at a temporal distance would conflict with basic assumptions about the physical world: this is of course conceivable (quantum non-locality, for example, might turn out to have applications to psychophysics), but would be an attack from within another powerful theoretical and scientific framework. Assumptions are rarely ditched without replacement, and purely negative complaints are doomed to be ignored. The positivist prudery which prohibits the postulation of intervening unobservables is one temperamental preference (Hesse 1961/1970: 291–2), and ‘suspicion of the hypothetical’ can play a useful whistle-blowing role in curbing scientistic excess: but investigation of processes and mechanisms underpinning capacities will not just stop.

Indeed, as Frank Jackson points out to me, there is a specific difficulty in postulating action at a temporal distance in the case of memory. Without causal continuity, a past event would not only have to leap to the present to cause my

8 For there may, for instance, be good empirical reasons why empirical research into memory traces is very hard. For sensible remarks about why, within the context of the history of the brain sciences, it should have taken so long to get any grip on identifying the trace, see Horn 1985: ch. 1 and Rose 1993.

9 In The Analysis of Mind Russell suggested that there is no a priori objection to the immediate causation of present events by past events, even when this means that ‘part of the cause has ceased to exist’ (1921: 89). A special, unique mnemonic causation might operate directly over remote times. But Russell is only arguing against the logical necessity of traces: while, he wrote, ‘in the present state of physiology, the introduction of the engram does not serve to simplify the account of mnemonic phenomena’ (1921: 85), there is no suggestion here that a full positive scientific theory of memory which did postulate traces would be conceptually incoherent. Russell’s words have been used to encourage acceptance of the possibility of action at a temporal distance in remembering by Anscombe (1974/1981: 126–7, in a defence of memory as ‘an original phenomenon of causation’), and Heil (1978: 66–9; compare Beardsmore 1989). But the bare possibility that mnemonic causation might be completely unlike any other causation is no reason to think that it actually might be. For defences of trace theory on this count see Broad 1925: 456–60, Warnock 1987: 46–52. Shope (1973: 318–21) is hostile to the use of traces in an analysis of remembering, but acknowledges that a developing scientific theory of traces could generate support for a causal theory of memory. Shope’s defence of mnemonic action at a temporal distance is mild, boiling down to a preference for the ‘economy of postulating the existence of fewer relations of causal relevance among phenomena’ (1973: 321).
memory activity; further, it would somehow have to track my spatiotemporal path, to ensure that it could at any time become causally active as I moved around. Such long-distance tracking of rememberer by past event seems unlikely to be ‘direct’ in any intuitive sense.

So another way of making space for traces while retaining direct-realist intuitions is by asking how in remembering I am in contact with the remote past. The problem is reminiscent of the ‘time-lag’ argument in the perceptual case. Events we perceive might have occurred before we perceive them, and the states of objects which we perceive might be temporally prior to their states at the exact time we perceive them. Orion, as we see it, is centuries old (Laird 1920: 45). Sound and light waves from the external world take time to reach our sensory transducers. Does the temporal gap between the occurrence of thunder and the hearing of the thunder not entail that perception of thunder is indirect, mediated, or inferred from the present stimuli? Not necessarily, for anti-representationists assert that we’re just directly aware of something existing or occurring at a time earlier than the present (Woozley 1949: 55–8; Heil 1983: 128; Ben-Zeev 1986: 299; Snowdon 1992: 76–7). This response, however, seems harder to maintain in the case of memory. It is not just that the temporal gulf in memory can be much greater than the barely perceptible gap between a distant axe striking its target and the sound of the blow reaching our ears: it is that the causal connections between remote past and current remembering are often more devious and twisted, enmeshed in more mixed causal fields and passing through more media, than the simple transmission of light or sound over time through a single medium. Some Wittgensteinian approaches, as I have noted, hide or deny the apparent temporal gap to be bridged between past experience or present remembering (Malcolm 1963: 238). It is hard to find clear alternative accounts: for John Laird, in remembering certain ‘peaks in the past’ do ‘stand out in relief where everything else is a bland or a crapulous haze’. This is metaphor, as Laird acknowledges, and ‘falls short of demonstration’ (1920: 52–3). Even if ‘direct acquaintance’ with what is temporally remote is a human capacity, questions remain about how it occurs. There is no obvious reason why memory traces might not be just what underpin such capacities.

10 Malcolm claims to have disproved all candidates like the trace which might have filled such a gap, and concludes that ‘in a sense, therefore, we do not know what it means to speak of a gap here’ (1963: 238). Warnock (1987: 45) comments that ‘in spite of the authoritative echoes of Wittgenstein . . . what Malcolm says is great rubbish’, because the temporal gap between the past and the present ‘is the most obvious gap in the world’.

11 Shannon argues (1991: 92) that the bizarreness of action at a distance is attenuated by attention to the context of action in the world, in which memory is used as a temporal tool with which to gain access to remote events. But, while acknowledging that, on this view, ‘memory is a process that enables one to remember’, he suggests that how this is achieved may not be ‘the subject of psychological theory at all’. The ecological wishes are laudable: but legislation over disciplinary boundaries is unlikely to succeed, certainly without support from a principled defence of the autonomy of psychology from other sciences.
16.3 Retention and storage

Critics of the spatial metaphor and traces claim that the retention involved in memory (whether this is the retention of knowledge, beliefs, dispositions, abilities, or whatever\(^{12}\)) does not require any physical storage or causal continuity across time. Therefore the postulation of traces or representations spatially located in the brain between past experience and present remembering, is unwarranted speculation and should be dropped (Squires 1969; Malcolm 1977; Schumacher 1975/6; Lewis 1983).

Malcolm notes, for example, that people can retain their good looks without keeping the good looks in storage. I can retain a sunny disposition after hardship without my disposition being stored (1977: 197). The fact that indigo curtains have kept their colour does not entail any causal connection between their previous states and their present state (Squires 1969: 178). So while memories, like beliefs, have duration in time, duration does not imply storage: ‘retention and storage are different concepts ... Storage implies retention, but retention does not imply storage’ (Malcolm 1977: 197–8). The confused identification of retention with storage mars not just philosophical but also neuropsychological writing on memory, which has created a ‘mythology of traces’ (Malcolm 1977: 199ff.). Again, think (says Squires) of a teddy bear in the attic which has retained its squeak. Even though six months ago, in the damp winter attic, the bear remained silent when poked, now that it is summer again the bear responds. This case of the retention of the ability to squeak, on Squires’ view, proves that no continuous causal chain over time is required for retention. ‘It is usually incorrect to claim that a toy that has kept its squeak squeaks now because it could squeak before’ (Squires 1969: 181). Memories, analogously, can be kept without a trace (Squires 1969: 196).

These critics feel something magical in the notion of causal processes operating over time (Bursen 1978: 19–47). A straightforward response for the trace theorist is to point to the permeation of ordinary thought and language by assumptions of causal continuity. Causal talk is restricted neither by requirements that instances of causation must be covered by general causal laws, nor by the specific metaphor of the serial causal chain from which Squires wanted to unchain memory. Talk of the ‘sources’ of a person’s present memory knowledge presumes some kind of historical continuity (Warnock 1987: 50–2). Normally, the precise nature of the continuity is unknown, for, as Deutscher says (1989: 60), ‘common sense and the thought expressed in common language requires only that there be some such [underlying] process’. The trace theorist does not have to say, as Squires seemed to think, that it is the retention

---

\(^{12}\) I cannot go into conceptual debates over whether memory is adequately analysed as the retention of knowledge, and whether it can also be a source of knowledge: see Ryle 1949/1963: 257–63; von Leyden 1981: chs. 1–2.
of the indigo colour by the curtains yesterday which 'causes' them to be indigo today. This is ludicrous (and Deutscher is 'happy to join in the joke'): and yet there are causal explanations for their retention of colour, to do perhaps with the effects of sunlight on particular dyes. The relevant causal conditions, those which explain the present colour of the curtains, the present squeakiness of the bear, or a present act of remembering, can be made more complex and inclusive, to cover, for instance, the causal effects of the environment over time, the teddy bear's causal field.

Behind this line of attack on causal theories is the view, associated with Ryle, that the dispositions which support counterfactual conditionals are pure powers, not grounded in any physical facts which make them true. There are many more or less respectable positions on the metaphysics of dispositions, and I only want here to point to the fact that there are views of dispositions which will allow explanations of memory capacities in terms of the nature of underlying states and processes. It does seem odd, even within common-sense understanding, 'to fancy that the retentive power of memory consists in nothing and subsists as "pure power"' (Deutscher 1989: 60).

But, even if room can be made for the quest for the exact nature of the operative involvement of the trace in remembering, and if some such criteria can be made to work along Martin and Deutscher's (1966) lines, this only goes to defend a very general causal account of memory. Yet there is no general issue of the links between retention and storage: it all depends on what it is that is stored or retained. So work still has to be done in spelling out just what is meant by saying that traces are 'stored' between experience and remembering. There are forms of (literal) storage which are not implied by retention: but Malcolm was mistaken in thinking retention in memory is possible without any causal continuity.

This is where the specifically distributed model of memory helps. If remembering involves the retention of capacities over time, does it require brute physical storage? The distinction within distributed models, familiar by now between explicit and implicit representation (chapter 6; O'Brien 1993) operates here.

The point of the distinction between transient activity patterns and enduring dispositional traces is that there is no need for representation which is both

---

13 As Deutscher points out (1989: 60), while the details of the chemical processes involved are the province of experts, the existence of such processes is commonly accepted, since, for example, 'knowing that dyes are complex chemicals is part of the everyday practice of following recipes'.

14 'To possess a dispositional property is not to be in a particular state, or to undergo a particular change; it is to be bound or liable to be in a particular state, or to undergo a particular change, when a particular condition is realized' (Ryle 1949/1963: 43).

continuous and explicit, for memories retaining their form in local boxes, as in
cruder causal theories of memory (the teddy bear is not always squeaking).
Context, in the form of the effects of past experience and new input on connec-
tion weights, automatically influences the transient explicit activation pattern
which is evoked on presentation of a new input. What is retained and, implicit-
lly, 'stored' in weights or pores after experience is a disposition. The actualisa-
tion, or rendering explicit, of an activity pattern at a given time depends on the
global state of the system, the individual connection weights, the present pat-
terns of input to the system, and thus on contextual environmental factors like
winter damp in the teddy bear's attic.

So it is true that retention in memory does not require the literal, faithful
storage of past experiences such that, if you open a door, they are just there.
Patterns which are not active exist only potentially in the changed weights of the
system. Remembering of course depends on many factors other than the
nature of what is 'stored'. But retention in memory requires some theoretically
explicable causal continuity over time through physical processes in which the
system takes part. The processes underlying our capacities may be much more
chaotic than any that sets of stored copies could pursue. Intuitions not hooked
on high fidelity or on the paradigms of print will cleave to traces which
inevitably mingle. No critique of trace theory can legitimately fail to deal with
implicit distributed traces.

16.4 Isomorphism and resemblance
How is it that traces represent past events, experiences, or whatever? What is it
about these 'representations' that allows them to latch on to the world and the
past? How can they have content, and what kind of content is it?

No philosophy of mind has a fully satisfactory account of the meaning of
mental representations. But specific problems seem to arise for the kinds of
mechanistic causal models I have described. There seems no coherent way to
spell out the required structural resemblance between the trace and what it is a
trace of. This resemblance relation between experienced order in space and
time and functional order in correlated mind/brain processes rests on
unanalysed and incoherent metaphors of imprinting, engraving, copying,
coding, or writing (felicitously dubbed typography, iconography, and engram-
matology by Krell 1990: 3–7). What does it mean to say that a trace is 'similar to'
that of which it is a trace?

Some critics focus on the assumption that the world, the past, or experience
has a structure to be somehow encoded into the brain. They claim, on the
contrary, that the world has no single, natural, non-arbitrary structure of ele-
ments to stand in isomorphic relations to internal traces (Malcolm 1977: ch. 10;
the idea that there are simple constituent parts of a chair). But the notion of the trace as a ‘structural analogue’ is sufficiently unclear to question without having to suggest in turn that all categorisation is arbitrary.

Neither do critics need to rely on assimilating all representations to images which picture the past: it is misguided to announce that trace theorists’ materialistic ‘dogma makes the copy theory indispensable’ (Straus 1970: 54; compare Locke 1971: ch. 1; and Hass 1991). ‘The traditional attraction with the image as the nerve of recall’ (Deutscher 1989: 71) only confuses the issue by demanding representation by resemblance.\(^{16}\) It is implausible that we carry round whole sets of mental pictures in our heads. Even apart from the difficulties of representing abstractions by resemblance, the mental image in fact sits badly with even a moderate physicalism. Brain states can hardly be similar to furry tabby cats or big rubber balls (Cummins 1989: 31–2).

So weaker notions of the trace are sought. Martin and Deutscher (1966: 189, 191) included in their analysis of remembering the requirement that, to the extent remembering succeeds, ‘the state or set of states produced by the past experience must constitute a structural analogue of the thing remembered’. Only this condition, they felt, would get around problem cases in which, for instance, present knowledge of something past is caused by a route other than memory. Martin and Deutscher did not want to commit themselves to the trace as a perfect analogue, ‘mirroring all the features of a thing’, or to there being any ‘sense in the notion of all the features of anything’: but, they believed, the trace must contain ‘at least as many features as there are details which a given person can relate about something he has experienced’ (1966: 190).

Even this now looks too strong. Some details which crop up in remembering an experience need not have been permanently encoded in the same trace as that experience. We tell more than we remember: later inferences creep into the telling. Even where memory for the gist of an event is roughly accurate, details may shift as the trace is filtered through other beliefs, dreams, fears, or wishes (Schacter 1996: 101–13). Less accurate verbal memories, such as attempted descriptions of faces or of wines, can overshadow more precise non-verbal traces (Schooler and Engstler-Schooler 1990). No psychological evidence suggests that remembering can be explained without causal connections between events and traces, and between traces and recollection: but it does encourage us to think of those causal connections as multiple, indirect, and context-dependent. So details remembered need not have been contained in the same trace in unchanging form.

Representation by explicit resemblance has often been specifically denied by

\(^{16}\) Malcolm says, rudely, that although the terms image, picture, idea, copy and representation ‘all tend to be equivalent terms in the philosophy of memory’, most memory theorists prefer ‘representation . . . , since its greater vagueness gives [their] thesis better protection against refutation’ (1970: 67).
trace theorists. Descartes rejected ‘little pictures formed in our heads’, which would require there ‘to be yet other eyes in our brain’, with which to perceive them (Dioptrics IV, AT vi.112, CSM 1.165; Dioptrics VI, AT vi.130, CSM 1.167). He noticed instead the significatory utility of things like words and signs, engravings, laughter and tears, which ‘bear no resemblance to the things they signify’ (Le Monde 1, AT xi.3–4, CSM 1.81; see also Principles IV.189–98, AT viii(a). 315–23, CSM 1.279–85; compare Gaukroger 1990: 14–26; and Slezak 1990). Brain traces are the vehicles of representation, but they require a causal nexus involving body and environment as their background system (Mackenzie 1989). The only ‘resemblance’ in play is between different traces, with some grouped together by, for example, being located in proximity in an internal state space, not between the structure of traces and the structure of the world: yet the world’s structure is still part cause of the structure of the traces.

Some contemporary critics of representation-by-resemblance who wish to retain representation instead follow Fodor’s account of the discrete storage of abstract symbols in a language-like representational medium (Pylyshyn 1984). But pictorialism and logicism do not exhaust possible forms of representation (compare Tye 1991): distributed representation could be a distinct third genus, neither image nor symbol (Haugeland 1991; van Gelder 1991b). Computations are defined over dynamic activation patterns rather than icons or syntactic strings (Kosslyn and Hatfield 1984; Hatfield 1989; Horgan and Tienson 1989; Cussins 1990). For Hatfield, for example, the relevant kind of structural isomorphism is simply the complex physiological transduction of, say, visual information into the visual system. There’s obviously immense difficulty in forcing a theory of content from such alternatives: But here I only need claim that metaphors of writing, imprinting, and copying do not exhaust the ways of getting isomorphism between past events and experience and encoded trace. If

17 ‘An activation space will also be a similarity space, inasmuch as similar vectors will define adjacent regions in space. This means that similarity between objects represented can be reflected by similarity in their representations, that is, proximity of positions in activation space. Similarity in representations is thus not an accidental feature, but an intrinsic and systematic feature’ (Churchland and Sejnowski 1992: 169). This is not a simple two-term resembling relation between image and object, but a four-term relation in which the similarity between two traces maps non-accidentally on to similarity between two objects or events.

18 The biggest new problem with this kind of approach (which looks like the right problem in this area) is in finding a principled way to distinguish which of all the subsequent activity patterns in the system are truly psychological visual memories, truly representational states, and which are mere physiological responses, merely causal/correlational relations between world and brain. It is for lack of a solution to this difficulty that Hatfield rejects the Churchlands’ notion of calibrational content, and prefers to tie representational content to the functional role played by the activity patterns in the system, and to those theories of content which make reference to the biological idea of normal function (1989: 269–76, 284 n. 41). But it is not altogether clear that a principled distinction between psychophysiological states and merely physiological states is necessary for good theory (Schwartz 1994).
there is any hope of more strictly neurophilosophical accounts of the isomorphism of patterns then some criticisms of trace theory may not get a grip and the notion of a structural analogue can be successfully generalised or weakened. To note that the structures which underpin retention need not remain the same over time, or might not even involve determinate forms over time, is not to refute the very concept of a trace. ¹⁹

16.5 The four-pronged fork: homunculus/regress/circularity/solipsism

Thus far I have tried to answer or deflect various hostilities to the trace. But a fourth set of objections, synthesised from the critical literature on memory and perception, has direct reflections in different options and possibilities within connectionist theory.

The four prongs of the fork which I will describe are not mutually exclusive: according to many critics of representationism, many trace theorists are confusedly pinioned on all four simultaneously. It is convenient to present the problems in the context of the retrieval process. How does the stored trace of the past come to play a part in the present act of recognition or recall?

Encoded traces, it is said, require an internal interpreter or reader to recognise a new input as matching an existing trace, or to know in advance which trace to search for and recall for a given purpose. But who is this inner subject behind the engram? Such an intelligent homunculus, merely shifting the problems of retrieval one step deeper inside, is the first prong of the four-pronged fork (Gibson 1970/1982: 95, 1979a: 255).

If it is avoided by having further internal mechanisms, in some 'corporeal studio' (Ryle 1949/1963: 36), pick out the relevant trace, a vicious Rylean regress, the second prong, is generated: each of these mechanisms will need a further mechanism to do its interpretative job for it. The problem of representation has been pushed further inside, not clarified.

Or if the homunculus is avoided by allowing that the remembering subject can somehow just choose the right trace, or match a new input to an existing encoded trace, then the trace would be redundant, for the subject has direct independent access to the past to know what a trace is a trace of. If the subject can just inspect the past in order to check and compare a representation against it,
the facts which traces were meant to explain have not been explained, and trace theory is circular (Earle 1956/7: 5–6; Buresen 1978: 52–60; Lyons 1981: 152–3; Wilcox and Katz 1981a: 229–32; Sanders 1985: 508–10; Arcaya 1989: 102–5; Krell 1990: 17–18, 89). This redundancy or circularity objection is the third prong of the fork.

If, finally, it is claimed that no independent access to the past is in fact assumed by representationism, and that no interpretations are being checked against another somehow guaranteed veridical memory, then, the critic argues, solipsism or scepticism results. There is no guarantee that any act of remembering does provide access to the past at all: remembering is all inference, and nothing gives us confidence in its accuracy. Representationism, it seems, cuts the subject off from the past behind a murky veil of traces (Wilcox and Katz 1981a: 231; Ben-Zeev 1986: 296; Maze 1991).

Handling the fork

These four prongs, homunculus, regress, circularity, and solipsism/scepticism await the unwary trace theorist. Having the recognition of traces done by a little internal recogniser merely pushes the problem inside, and homunculus or regress arguments ensue. But if such inner recognition does not occur, then either the subject has circular independent access to the past to compare with the trace, or there is no guarantee of any access to the real past, and solipsism threatens.

To some extent these repeated worries are just ignored by memory theorists in the cognitive sciences, who continue to seek existence proofs that remembering is possible with representationist assumptions. Defences depend on the particular version of representationism being fortified (Shoemaker 1970; Rakover 1983; Rock 1991). Localist models of memory like Hooke’s (chapter 5 above), with atomic items independently stored at separate addresses, require a central processor to extract and manipulate them, an executive Chairperson of the Board of Mental Modules (Krell 1990: 89), and some defend this homunculus (Rock 1983: 39; Fodor 1986: 320–1). But I do not pursue this defence, since central executives are exactly what I hope to dispense with.

Others deny the force of the regress prong, claiming that the ability to ‘interpret’ sequences of on- and off-currents can be decomposed by subdividing the tasks involved and attributing each to an ever smaller ‘team or committee of relatively ignorant, narrow-minded, blind homunculi to produce the intelligent behavior of the whole’ (Dennett 1978: 123) and that there is no illegitimate intelligence in mere mechanisms (Fodor 1968b: 627–30).²⁰ This

²⁰ As Dennett argues, ‘homunculi are bogeymen only if they duplicate entire the talents they are rung in to explain’, which is not the case when the nested levels of homuncular function are decomposable to, for instance, simple stupid binary mechanisms: ‘one discharges fancy homunculi from one’s scheme by organizing armies of such idiots to do the work’ (Dennett 1978: 124).
decompositional strategy for innocent homuncularism has been connected with recent work on function-analytical explanation to accommodate connectionism (Lycan 1991). It has attractive consequences in stressing the continuity of levels of nature which rejects the simple 'two-levelism' of earlier philosophical functionalism (Lycan 1987: 37–48, 1991: 265–8). But while it may be that biting the homunculus and regress prongs will eventually blunt the four-pronged fork, it is not immediately obvious that functional decompositions explain how stored traces get matched to inputs or how retrieval in general works.

Perhaps a more promising response is to deny the force of the final, solipsistic/sceptical prong of the fork. In distributed models, there is no internal central processor which searches, inspects, or otherwise manipulates items stored passively in discrete memory locations. The only connections within the system are local, and processing and storage occur in the same parts of the system. There is no internal reading, decoding, or interpreting of memory traces. Behaviour mediated by cognitive activity is the result not of an intelligent homunculus' calculations, but of numerous relatively independent but interacting systems computing best-fit solutions in parallel. But, without a homunculus, how do recognition and learning happen? Can the past in fact play any role in the causation of present retrieval episodes? In the context of the right kind of causal theory, the world itself must partly drive the learning and remembering process.

So the direct realist's circularity charge contains an element of truth important for the connectionist. As Gibsonians argue, information about the past is in the present environment for the attuned subject to extract. The past is not forever concealed: sequential orders in time allow veridical remembering on the prompting of present inputs. The connectionist is enough of a representationalist to think that accepting the unique specification of the past by present input alone would be to ignore the internal and cognitive effects of past experience (chapter 15 above). The same present input or cue will not produce the.

21 Compare Fodor 1976: 204–5 on how, as a representationalist, to avoid thinking of 'the mind as somehow caged in a shadow show of representations unable, in the nature of the thing, to get in contact with the world outside'. The causal nature of the sequence of events from objects to representations cuts off a degeneration into solipsism, so that 'epistemic relations are typically immediate in whatever sense causal relations are, and that ought to be immediate enough for anybody'. But Fodor's representational theory, attributing much to innate mental structure, does not say more about the role of the world in the formation of representations: the connectionist response which I am adumbrating does. It also deflects John Maze's case (1991: 164–8) that Fodor's representationism cannot avoid the 'decline into solipsism'. Maze (1991: 167–8) complains that, while Fodor (1976: 204) is right in saying that solipsism cannot be true because representations must have causes, his representationism does not allow him to talk about these causes. This is not true of the representationism I am defending, according to which it is not that we are only ever aware of representations, but that we can sometimes be aware of the actual external causes of the representations.
same response in two different rememberers with different learning histories, or in one rememberer at different times or in different contexts. Input comes from elsewhere within a system or organism as well as from the environment. But while the connectionist does take memory to involve inference in an important way, realism is maintained by arguing that the ‘teacher’ of the connectionist model or mind is the environment, plus evolution. This needs some cashing out.

**Circularity and supervised learning**

The final two prongs of the philosophical fork, the circularity and solipsism objections, have direct empirical realisations in the options available within connectionist theory. ‘Supervised’ learning rules require circular independent access to reality, the past, or the right answer, making the postulation of representations redundant. ‘Unsupervised’ learning rules, on the other hand, render the system’s contact with the world problematic.

Connectionist models employ a variety of learning rules (Hinton 1989; Bechtel and Abrahamsen 1991: ch. 3). These are algorithms by which the weights on connections between units change over time. The error margin, in a simple case, between an actual output and the ‘correct’ output can be used over a series of trials to alter the weights on the units which together compute the output value. Such supervised learning rules, like the delta rule, back-propagation, and Boltzmann learning with simulated annealing (Churchland and Sejnowski 1992: 82–96) are common, since they allow easier management of ongoing processing: ‘the network is provided with explicit feedback as to what output pattern was desired for a particular input pattern’, and desired output is compared to actual output in order to change the weights (Bechtel and Abrahamsen 1991: 314).

The importance of supervised learning in human development should not be underestimated: as well as learning by ostension and schooling, explicit correction by external teachers is one mechanism for the internalisation of a culture’s social and moral categories. But supervised-learning techniques have major shortcomings if considered as complete models of human cognition. Paul Churchland, writing on back-propagation in particular, notes:

> First, the requirement that the correct output be available to the learning network in every case is clearly unrealistic. After failing to solve a problem, real brains do not generally get to look at the correct answers at the back of the book. Second, the brain shows no plausible mechanisms for computing and

---

22 What seems a really robust realism is the possibility that we could be globally mistaken about the way that the world (or the past) is. Global realism must allow for large-scale error (Churchland 1988; Pettit 1991). While, I shall argue, scepticism does not follow from this, Gibsonian ‘direct’ realism does not seem to even allow for the possibility of significant error (see note 7 to chapter 15 above).
distributing such globally informed adjustments to its myriad weights. And third, the back-propagation procedure scales upward to large networks only very poorly. (Churchland 1989a: 243-4)

Churchland’s first complaint about supervised learning is the empirical version of the circularity prong. As he notes, we need ‘to escape the unreality of an omniscient teacher’ (1989a: 246).

The biological implausibility of supervised learning realises in empirical form the circularity or redundancy prong of the direct realist’s four-pronged fork. A system outside the system of traces already knows the answer, identifying whether a candidate stored trace is the appropriate one for current input. In human remembering, then, these models would be circular, for they require external access to the past to perform their retrieval operations.

**Solipsism and unsupervised learning**

Just as the circularity prong of the philosophers’ fork has an empirical realisation in connectionist research, so too does the solipsism prong, on which impaling must be risked. ‘Unsupervised learning’ labels a range of biologically more plausible learning algorithms. Without an internal interpreter, and without independent access to the past, neural networks must somehow maintain genuine causal contact with the world. This was the perceived problem with animal spirits models of memory, for the fleeting spirits were pre-eminently unsupervised, and threatened to blunt confidence in the possibility of veridical recall. The key, then, is the input. In unsupervised learning, networks ‘evolve processing strategies that . . . maximise their capacity for identifying salient information in the . . . input vectors . . . [and] find similarities among the inputs so that they are taxonomised into potentially useful groupings’ (Churchland 1989a: 246).23 The partitioning of hidden units groups inputs into similarity classes. The network becomes increasingly sensitive to complex regularities, through a ‘progressive accommodation to the objective statistical distribution among inputs’ (Churchland 1989a: 248).24 If something like this

---

23 Bechtel and Abrahamsen (1991: 314) describe the way a network using unsupervised learning ‘gradually achieves, without feedback, a weight matrix that allows it to classify a set of inputs (by discovering the regularities exhibited by subsets of the input patterns)’. The terminology here needs some care or modification, as P.S. Churchland and Sejnowski have pointed out (1992: 96–7): in unsupervised learning a network has no access to external feedback as from a programmer, but it can have internal error messages either from other parts of the system or in using its own previous states retained in some form in extra context layers (compare Jordan and Jacobs 1992). They suggest the term ‘monitored’ learning when there is some internal measure of error. For the philosophical issue at hand, unsupervised systems face the solipsism charge whether or not they are monitored: only systems with supervised learning have independent access to the answer.

24 It is for this reason that connectionist models promise better performance than expert systems over some complex task domains. Human experts’ skills need not be accessible to conscious, linguistic or explicit formulation (Dreyfus 1990). Nevertheless skill is no accident: experts respond to objective regularities in the world.
occurs in biological networks, the connectionist has some hope of blunting the solipsist/sceptical prong of the direct realists’ dilemma, since the world itself can ‘exert a robust influence on the process’ (1989a: 248; P.S. Churchland and Sejnowski 1992: 202–21).

It is worth spelling out the philosophical dialectic here. The charge was that no representationist theory which avoids the other three prongs can avoid solipsism. The critics are not sceptics or solipsists themselves, and no ultimate refutation of solipsism is specifically required in answer. There is room for an empirical response which shows that in fact particular representationist trace theories do allow for genuine contact with the past: if this is so, the objection that the theory does not allow reality to impinge on the rememberer loses its grip.

But not all realist critics are content with the fallible, corrigible processes which such empirical answers provide. There is a tendency to demand not probable causal contact with reality, but certainty in remembering as in perceiving, incorrigible contact with the world. Turvey and Shaw, for example, complain that ‘the muddle’ arising in perceptual representationism is due to ‘a failure to uncover ... any legitimate kind of knowing that is epistemologically unquestionable’ (1979: 178). They acknowledge the suggestion, from Fodor and others, ‘that such epistemic incorrigibility is not really needed’ and that there need merely be some causal process or ‘algorithm’ which in fact underpins contact with the world although it is not ultimately ‘unquestionable’. They reject this solution because it ‘bears the traces of legerdemain; an unquestioned but incorrect algorithm as a bench mark could hardly guarantee that other algorithms will be adjusted in the direction of a truer fit to reality’ (1979: 178). But the demand for a guarantee of incorrigible access to reality is terribly strong foundationalism: as Schwartz (1996: 89) complains, it is hard to be convinced by the Gibsonian view that ‘unless there is a one–one, lawlike, correlation between the environment and some single aspect of the stimulus array perception would be indirect’ in a metaphysically troubling way. The imposition of certainty should be rejected. Why look beyond the empirical likelihood, given natural selection, that contact with an objective world does, sometimes, occur? The cost of a robust realism just is the bare logical possibility of scepticism. But bare logical possibility is insufficient ground for concern or action.

A fallibilist defence of the reality of the external causes of representations is hostage to the empirical success of the theory in which it is embedded. Neurobiological research on long-term potentiation may show how unsupervised learning is realised in some brain regions (Goddard 1980; Lynch 1986; Sejnowski and Tesauro 1990; Alkon 1992): connectivity is adjusted locally at independent synaptic sites without a global, omniscient teacher. The different mechanisms involved at different temporal stages of consolidation are driven only by environmental effects and the current state of brain and body, arising
from evolution and past experience (Thorpe and Imbert 1989; Churchland and Sejnowski 1992: 254–95).

Distributed models extract information from inputs, becoming attuned, in context-dependent fashion, to what the environment affords. Despite the promissory tone of this discussion, there are good philosophical reasons to look in these directions. We attune ourselves to our environments, sometimes through explicit tuition but more often by picking up patterns in experience, revising and transforming internal models as we go. It would be strange if empirical theories described the mind/brain as faithfully reflecting the world in its full presence, or the past in its full transparency, as the demand for epistemologically unquestionable remembering requires. Better metaphors are those of the continual filtering, deformation, and melding of representations over time (Churchland 1993: 220–1). Body, brain, and memory do actively alter perceptual input, not by a dualistic inferential imposition of constructed or innate structure on impoverished data, but through ongoing rearrangement and moulding of representations in the interaction with, and 'interpretation' of, new experience.

To put it another way, of course truth in memory is a problem. Just as Descartes and Malebranche realised that the same processes operate in animal spirits across memory, imagination, and dreaming (chapters 3 and 9 above), so connectionism allows multiple causal inputs in any act of remembering. There is seldom a simple, direct transmission from a single past experience through discretely stored informational items to a cleanly defined moment of recall, for 'each memory is many memories' (Matsuda 1993: 24; Schechtman 1994: 9–10). Outside philosophy and the courtroom, we only recognise human memory as operating 'normally' when it is 'malfunctioning', shot through with forgetting, selection, condensation, interference, and distortion. Yet, outside contexts in which truth is so desired that it is enforced, this rarely elicits worries about solipsism. The experienced sedimentations of memory in the body and of affect in memories make it blindly obvious that the past, for all its obscurity and opacity to conscious capture, does affect the present in multiple, skewed ways.