I

Introduction: traces, brains, and history

... it can easily happen that we cannot simultaneously store many ... structures in a connectionist memory without getting intrusions of undesired memories during the retrieval of a given memory ... (Paul Smolensky 1991: 218)

... how much similarity must there be between the two moments in order for the one to count as a memory of the other? How much of the content of the experience must be reproduced and how accurately? How many portions of the past is the present connected to in a condensed memory, and how is this determined? (Marya Schechtman 1994: 9–10)

1.1 Porous memory

Beyond the archive

Porous memories fuse and interpenetrate. Fragments of song mingle in hot remembered afternoons, mysterious angers return at a flush with a chance forgotten postcard. Such memories were once the motions of old fluids, animal spirits which meandered and rummaged through the pores of the brain. They held experience and history in bodies which were themselves porous, uncertainly coupled across tissues and skin with their air, their ethics, their land. Now they are patterns of activation across vast neural networks, condensing and compressing innumerable possible trajectories into the particular vectors of flashing or torpid memories. Dynamic cognitive systems coevolving with the physiological, environmental, and social systems in which they are embedded (van Gelder and Port 1995: 27–30) need the wishful mixings of absence which interfering traces bring.

These studies in the history of theories of memory are grounded in new interpretations of strange, neglected old French and English neurophilosophy. But only late twentieth-century worries about memory, science, and truth make sense of indulgent attention to ‘seventeenth-century French connectionism’ (Diamond 1969), and to bizarre historical beliefs about interactive relations between self, body, mind, and coursing nervous fluids. This kind of historical cognitive science aims to demonstrate that it is possible to attend to contexts and to brains at once.

It is no big deal now to claim that human memory is not a set of static records in cold storage, that the subtle smack of the organic opens remembering to decay and confusion, affect-ridden association, the pains of time. Not all theories have taken memory to be a place where dead parts of the past sit passive
INTRODUCTION: TRACES, BRAINS, AND HISTORY

until recalled to full presence. But, across the bewildering range of disciplines in which models of memory are constructed and criticised, vast gulfs between brains and society (felt even, or especially, by those who deny them) limit moves beyond the archive. The fact that, say, neurobiology and narrative theory, as well as cognitive psychology (Roediger 1996: 79), describe the constructive functions of errors or lapses in the fidelity of memory is not sheer accident. But, in the frantic rush of new research in all of memory's fields, it is impossible to consider the physiological, the cognitive, and the cultural at once. Old, rejected theories offer a feeling for the shape of some debates about control of the personal past which pre-date our debilitating, tedious battles between 'science' and 'humanism'. It would be nice to entwine philosophical, social, psychological, and neuroscientific accounts of memory in modern contexts alone, in wild anthropological fables about the phenomenology of neural nets: but the frameworks are still too disjointed, and so only history affords the requisite pretense of distance.

I undertake both the description and the defence of related theories of memory, from animal spirits to connectionism, which employ superpositional storage: memories are blended, not laid down independently once and for all, and are reconstructed rather than reproduced. In dissolving old and new lines of attack on such theories, I suggest that they exemplify the sensitivity to culture and history which good psychological science can exhibit. Working between historical and contemporary material suggests that wider issues about the self and psychological control are also implicated in current debates. The models of memory distributed through these studies, in mosaic from Descartes to connectionism, hint at a more reckless algebra, an understanding of how complex self-organising physical systems like us can be so psychologically plastic, attuned to the configurations of culture in which cognition and remembering are situated.

I cannot, of course, even begin to fulfil this promise: these studies are mere groundwork, trying to undermine various patterns of hostility to neurophilosophical theory. In too many spots I only sketch approaches to difficult puzzles, leaving detail undone. And yet there is a tenuous continuity between the studies, a faint order which might justify the threadbare juxtapositions. The active use of history in bringing culture into science and in undermining easy present-centredness requires a certain obliviousness, in theory as in practice. I hope that there is enough in these studies to excuse their shortcomings with respect to that relentless erudition which the genealogy of concepts and theories demands.

Interdisciplinarity
My account of these theories of memory both complicates and implicitly defends a set of philosophical positions crudely characterisable as meth-
anism, naturalism, associationism, determinism, and reductionism. These attitudes seem to signify thrall to matter, foolish scientism or misplaced physics-envy, blinkered materialism, the lack not so much of spiritual orientation as of embedding in culture. I do indulge overenthusiastic gestures and unlikely promises in these pages, through an untidy preference for proliferation over prudence in difficult domains. But I want to temper the repugnance which swells when wise humanists encounter cognitive sciences and neuromyths, by adding a sense of history, culture, and play to my reductionist neuro-philosophy. Amidst the vast literature on memory, specific and insistent interdisciplinarity aligns this book with other approaches, histories, and ideas which are not usually put together. Detailed historical analysis of theories of memory in medicine, neuroscience, and philosophy sits, at least, in unusual combination with gullible faith in the new sciences of complexity, memory, and brain.

Theories of memory are a test case for the wish to connect cognition and culture. In breaking down educational and cultural divides between arts and sciences, it must be possible to trace interactions between minds and their social surround, or between particular bodies and the worlds in which they grow. Even if the shared backgrounds and forms of life in which individuals develop can never be fully articulated, this means not that science or theory is restricted to the repeatable and isolable, barred from dealing with complexity and change, but that the social permeation of the psychological is the most puzzling and urgent of areas for attempts, at once scientific and cultural, at theory. Only thus can the sciences of the mind/brain ever usefully spill out of their institutional limits and tell those on the outside things they want to know.

So I seek to show how mechanists can also be holists, how determinists can also be contextualists, how naturalists can accept their engagement within frameworks, how bodies too can have narrative flows. Existing taxonomies of theories of memory (Belli 1986) are disrupted by these models of memory. The point is not just that science itself (as activity and as product) is in time and culture, but that it also comfortably deals with the time-bound and the context-dependent. Memory is both a natural and a human kind (Hacking 1994). Its operation, in species, society, or individual, does not alter easily, and it cannot be moulded at will, for the body and the past both resist arbitrary voluntary manipulation: but neither is it forever fixed, its processes or its contents shaped beyond change by preordained, pre-social forces. The various sciences of memory still display a puzzling lack of overlap (Hacking 1995: 199), and the one material world in which memories exist looks increasingly disunified and promiscuous. How in practice, in detail, do complexity and explanation co-exist?

To sceptics about the very idea of cognitive sciences, the 'memories' of our computers furnish only ludicrous analogies for human remembering.
The point of the information storage systems which permeate our life is to retain static items, unchanged unless manipulated. If brains and bodies are introduced, they are more likely as hardware than as wetware, as containers and conduits of independent information than as noisy or sedimented transformers. But anti-scientific zeal is too easily promoted by mocking reductionists as inconsistent every time they speak a language other than fundamental physics. For matter is in culture and time, nature is in history, the brains through which experience piles are not isolated. Memory bridges not just past and present, but outside and inside, machine and organism, dreams and reason, invention and sadness, creation and loss.

**Morals affect physiology**

And so the archive caricature of the cognitive scientific view of memory must be displayed, questioned, and lampooned. But challenges to rigid approaches to memory do not rule out all scientific study of remembering. Interference too has its patterns and constraints, confusion its formal operations. Clearer tracing of historical and contemporary debates reveals important distinctions not so much between scientific and non-scientific methods as between explanatory polarities of order and chaos, discipline and anarchy. Within scientific models the gulf between new connectionist and classical symbolic approaches to cognitive science is only the most recent manifestation of older divisions. Early modern moral physiologists did not need to abandon the discourses of natural philosophy or ‘science’ in order to make their recommendations on the pursuits of virtue and truth. So when I describe ‘tension’ between neurophilosophy and ethics, or show how physiological theories were revised to fit social demands, I am not enforcing a model of inevitable conflict in which the ‘scientific’ must pull against the normative. Unease about the body and the traces it conceals provoked crises within the best theoretical systems, for ‘knowledge’ of mind and brain often had to serve as both truth and morality (Smith 1992: 231–8). Interdisciplinarity is here easy to spot if difficult to carry off, for knowledge-that in theories of memory is always also knowledge-how, moral and practical at the same time as scientific.

For cognitive scientists, especially new connectionists, this embedding of mind, brain, and memory in body and culture is urgent. ‘Neurophilosophy’ (Churchland 1986a) may have sprung from frustration at philosophy of mind and from excitement at the wonders of computational neuroscience. But, despite critics’ laments at the ‘pervasive gloom’ of asocial materialist orthodoxy (Eccles 1994: x; Sharpe 1991), neurophilosophy would never work as ‘austere scientific abstract theory’ alone, and requires revisions of social, political, and historical understanding to run along with the revisionary philosophy
INTRODUCTION: TRACES, BRAINS, AND HISTORY

of psychology (Churchland 1993: 218–19, 1995: 286–94). Few have been both willing and equipped to embark on the task despite increasing recognition of its necessity (Hatfield 1988a: 732; van Gelder 1991a: 93). It is no use simply to complain that neurophilosophy is 'philosophically inadequate because it does not deal with the ethical dimension of the mind' (Stent 1990: 539, 556); but the new connectionist ethics being developed in response (Clark 1996) can be enriched with cultural and historical counter-theory to add to the brain-work and the morality. There is no moral theory in this book: but it does start to connect connectionists with dead revisionary allies and fellow wantons.

1.2 Distribution and dynamics
I invoke throughout a distinction or (better) a spectrum between local or archival models of memory as unchanging items in storage spaces, and distributed or reconstructive models of memory as blending patterns in shifting mixture. History and rhetoric pitch reproductive models of remembering against reconstructive, fidelity against fragility. Both old and new distributed models describe dynamic systems. Animal spirits theory, like some connectionist models, fits Tim van Gelder's description of a class of possible dynamical cognitive models, in which cognitive systems are 'complexes of continuous, simultaneous, and mutually determining change' (1995: 373):

the cognitive system is not just the encapsulated brain; rather, since the nervous system, body, and environment are all constantly changing and simultaneously influencing each other, the true cognitive system is a single unified system embracing all three . . . interaction between the inner and the outer is . . . a matter of coupling, such that both sets of processes continually influence each other's direction of change.

My attention, then, is on two versions of that subset of dynamic models which employ superpositional storage. In an appendix to this introductory chapter (pp. 19–20), I sketch the connectionist framework for readers unfamiliar with it. Here I introduce significant issues linked with the local/distributed distinction to explain why those outside cognitive science should care, then focus on this key notion of superposition.

1 Churchland and Sejnowski (1992: 445, n. 5) accept the importance of the social level for the neurophilosopher, while acknowledging that 'it has not been the main focus' of their book. The case for extending the relevant levels of research from synapses, networks, and maps to social interaction between organisms and their brains is that 'the interaction between brains is a major factor in what an individual brain can and does do'. I add that bringing in the social enriches neurophilosophy also by opening interaction with disciplines which start from the social, and thus newly moulding the explananda for a mature neurophilosophy. My project is to probe potential historical and theoretical advantages of some models of memory which allow for and invite such extensions.
Total recall

Surprising personal and social consequences flow quickly from unthinking acceptance of a local model of passive items in independent cells, splayed on the spirals of memory, at the beck and call of the executive individual who possesses them. I am unsure if the idea that all memories somewhere remain ordered and unblemished has ever been part of ‘folk psychology’ (it had, for example, to be enforced powerfully by the English Restoration philosophers I discuss in chapter 5). But British Telecom invest vast millions in a ‘Soul Catcher’ project, which aims at ‘memory transfer’ by picking out and playing back individual traces in another brain (Guardian, 18 July 1996, p. 1): this is not the gorgeous fantasy of interpersonal dreaming which drives Wim Wenders’ film Until the End of the World (1991), but a sad, expensive rerun of old ‘bizarre memory experiments’ which fed RNA from one worm or rat, in so-called ‘informational macromolecules’, to another so that the recipient could learn from the donor’s experience (Rose 1993: 189–99). Yet in one survey 84 per cent of psychologists and 69 per cent of others believed that ‘everything we learn is permanently stored in the mind’ and is potentially recoverable (Loftus and Loftus 1980: 430).

If atomic items did remain impermeable to further change after encoding, access to a desired memory in court or in therapy might be difficult, but would always be possible in principle. As both recovered-memory controversies and science fiction teach, the quest to reproduce the content of an original experience would often fail to comfort: the personal past would be tyrannical, events preserved in aspic always returning to haunt us (Spence 1988: 320–1). But whatever evidence of memory malleability, suggestibility, and distortion psychologists produce in response to moral panic about repressed memories of abuse (Loftus, Feldman, and Dashiell 1995; Schacter 1996: 248–79), it cannot be proven that some memories do not sit fixed in awful archives (Bowers and Farvolden 1996; Brewin 1996). But note also the immediate implication of views about the self in theories of memory. Local memories are kept in a storage system which is distinct from ongoing processing, in a dusty corner from which a possessive individual must try to remove them on request. Such a theory of memory is but a minor part of a theory of cognition, in which problem-solving and abstract reasoning can take precedence.

Distributed memories, in contrast, are troubling just because their content can change over time.2 If traces are composites, superimposed over long experience, what emerges in retrieval may be noisy, ambiguous, or systematically

---

2 Philosophers are sometimes sceptical about memory traces because, they realise, many factors other than brain states contribute to remembering. It is worth stating at the outset for their benefit that, obviously, theorists concerned with social aspects of memory must acknowledge that demands of specific situations affect the content as well as the expression of a memory. There is no reason to attribute to trace theorists the view that remembering is
INTRODUCTION: TRACES, BRAINS, AND HISTORY

distorted (Metcalfe Eich 1982: 611). Storage is naturally entwined with processing, and a theory of memory is central to a theory of mind. It is not that there are no secret angles of the mind, for in superpositional psychodynamics there is no easy conscious access to the forces driving representational change: but any ‘inner walls of secrecy’ where discontinuous systems coexist (Lingis 1994a: 148) in such models are immanent to the memory landscape, not imposed by executive decision.

As the anthropologists Michael Lambek and Paul Antze suggest, resistance to this idea that the sources of distortion may be internal and unavoidable is shared by those on both ‘sides’ of the false-recovered-memory controversies: ‘such is the need to shore up a space of organic innocence that its absence can only be imagined in terms of a deliberate and violent despoiling on the part of corrupt adults’ (1996: xxx n. 7). But they deny that the cognitive psychology of memory can help in encouraging acceptance of the complicated and inconsistent roles of remembering as a practice, on the ground that psychology inevitably constructs memory as ‘objective and objectified’ and omits ‘the relation to a self, agent, or community that bears memory’ (1996: xi–xii). In contrast, I show that cultural studies of memory too can find material of interest in dynamic models within connectionist cognitive science.

Confusion and mixture

On top of the basic, familiar connectionist propaganda outlined again in the appendix to this chapter, I examine more closely the central notion of superposition and its consequences. In true distributed models, memory traces are both extended and superposed, many traces piled or layered in the same physical system, with many ‘representations’ in one ‘representing’ (van Gelder 1991b, 1992a; Haugeland 1991; Schreter 1994). ‘Each memory trace is distributed over many different connections, and each connection participates in many different memory traces’: the traces of different memories ‘are therefore superimposed in the same set of weights’ (McClelland and Rumelhart 1986: 176). A trace is extended when it is spread across a number of elements or parts of a system, with many elements required for any one pattern. But extendedness is not enough for distribution, since every trace could still be quite distinct, entirely independent of the set of elements composing every other trace: such a model would still be local. Superposition, then, is also needed.

‘Two representations are superposed if the resources used to represent item 1 are coextensive with those used to represent item 2’ (Clark 1993: 17). Most determined by the properties of the stored item (compare chapter 16 below). Mainstream psychology deals in detail with factors other than the nature of the trace: research on Tulving’s ‘synergistic ecphory’ (1983: 12–14), for instance, describes the conspiratorial interaction of the cue (in the context of retrieval) with the trace (Schacter 1982: 181–9, 1996: 56–71). I am concerned primarily not with encoding or retrieval, or with cueing effects, but with alterations in traces during other ordinary ongoing processing.
distributed representations in practice are only partially superposed, on this
definition. Superposition gives traces an internal structure. Patterns of activity
grouped around a central prototype are subtly different from each other, but are
similar to varying degrees in various objective respects.

Many philosophies of mind have trouble dealing with ‘difference’: the
models they concoct reveal minds endlessly assimilating multiplicity to iden-
tity, turning threatening or challenging variation into safe and comprehensible
repetition. But distributed models, in contrast, have problems with sameness.\(^3\)
Public representations like sentences may be frozen, relatively memorable,
‘context-resistant’, and thus relatively stable (Clark 1997: 210; but compare
Sperber 1996: 25, 58, 100–6). But every occurrence of a mental representation
is different, because every explicit tokening of a pattern of activation is a
reconstruction: this leads connectionists, in the extreme, to say that we never
create the same concept twice (Barsalou 1988: 236–7; Clark 1993: 91–4).\(^4\) In
connectionism, says Elman (1993: 89),

\[
\text{once a given pattern has been processed and the network has been updated,}
\text{the data disappear. Their effect is immediate and results in a modification of}
\text{the knowledge state of the network. The data persist only implicitly by virtue}
\text{of the effect they have on what the network knows.}
\]

Thinking of mind as text, of mental representation as language-like, made it
easy to assume that sameness of meaning is unproblematically transferred
across contexts. Words, normally, retain their meanings across different
instantiations: ‘apple’ is easily recognisable whether scrawled misshapenly in
a recipe notebook or printed in neat Palatino font in crisp poetry. The difference
between tokens rarely challenges the sameness of type. The same information
can thus be drawn on in many different circumstances, and is multiply usable
without degradation. The point of language, or of a language of thought, is to

\(^3\) There is one sense in which this is an issue for any materialist theory of memory, given
the incessant motion of matter. Critics are led to reject physicalism: Straus (1970: 50)
notes that ‘in physics and physiology events are not repeated’, and concludes from a
phenomenological examination of memory that ‘experience, then, transcends the realm of
physical events’. But distributed models have specific problems with sameness, since
superposed traces have not even the kind of imperfect but enduring material continuity
possible for single stored items like books, bags, and birds.

\(^4\) This is the key sense of ‘reconstruction’ in my talk throughout of reconstructive memory.
It does not mean that the deliveries of memory are always false, or that the fragility of
remembering should override common-sense trust in memory or testimony: I do not want
to belabour ‘that banal topic, the indeterminacy of memory’ (Hacking 1995: 234).
Obviously, as Coady (1992: 268) observes about eyewitness testimony, ‘neither the picture
of wholly passive registration nor that of furiously active invention’ tells the complete
story. Rather the notion of reconstruction marks the content-addressable nature of
memories, and the context-constrained nature of every act of remembering: the extent to
which the picture of highly nuanced mental episodes, specifically indexed to the cognitive
system, body, history, and current cues in which they occur, is alien to ‘common sense’ is
But thinking of mind as process, with representations in distributed rather than linguistic form, means that current context is built into the particular reconstruction of any one pattern.

This feature arises directly from superposition. Since many traces are ‘stored’ in the same physical system, no single one of them can be continually explicitly active. Memory cannot be the permanent conservation of discrete unchanging informational atoms. But what then is the memory trace? Where does the trace disappear to between experience and recall, between past and present? There is only one set of connections in any system, only one set of weights between connections, while there are many traces. So traces are affected on reconstruction by the other traces implicitly present in the system, and may blend one with another, leading potentially to distortion or error.

Some connectionists try to exclude interference and the potential confusion between traces which it brings. Patterns of activity are set up to be independent enough to minimise blending between different traces encoded in the same representational resources: ‘if the patterns are sufficiently dissimilar (i.e. orthogonal), there is no interference between them at all. Increasing similarity leads to increased confusability during learning’ (McClelland and Rumelhart 1986: 185). A priori legislation against confusability, in favour of non-destructive overwriting (Tryon 1993: 344), is tempting: ‘By “superpositional storage” I mean the property that one network of units and connections may be used to store a number of representations, so long as they are sufficiently distinct (the term used is “orthogonal”) to coexist without confusion’ (Clark 1989: 100).

Motivation for thus excluding confusion from distributed representations comes from fear of ‘catastrophic interference’ when models are realistically scaled up (McCloskey and Cohen 1989; Ratcliff 1990). This occurs when the learning of a new set of data wipes out memory of previous data and successful reconstruction becomes impossible, when the mixture’s ingredients will not reseparate. ‘Catastrophic forgetting is a direct consequence of the overlap of distributed representations and can be reduced by reducing this overlap’ (French 1992: 366).

Unleashing interference
The specific sources of such disastrous interference in distributed models are disputed (Lewandowsky 1991). But neither this debate nor the slightly moral tone of some false-memory research on suggestion and misinformation illusions should blind us to the startling productive role of interference which fuelled connectionist enthusiasm as soon as data on interference in human memory (Anderson 1995: 247–65; Rubin 1995: 147–55) was modelled in neural nets. The same mechanisms which induce false recognition of plausible information (Roediger and McDermott 1995) also drive flexible generalisation
and the capacity to ‘extract the central tendencies of a set of experiences’ (McClelland and Rumelhart 1986: 193; Clark 1989: 99). Composite traces blur and fuzz the memories of specific episodes, but render salient the overlapping, prototypical features of a set of exemplars:

Each time an event occurs in a different context (time, place, and so on) a new trace is formed, but soon there are so many different contexts that none can individually be retrieved. What is common among the several exemplars is the knowledge, which we call abstract, but by default, by the massive interference attached to any individual context. (Crowder 1993: 156)

Even traumatic memories of repeated or persisting events may be filtered through later emotions: in such cases, memory is often accurate enough for the general character of the events, but awry in specific instances, mixing together the thoughts, perceptions, and emotions of different occasions (Schacter 1996: 205–12). It may be dangerous to unleash interference in contexts where historical truth matters terribly: but the fact that, like neural nets, humans often fail ‘to separate information that arises from different sources’ (McClelland 1995: 73) is also a powerful fund of pleasure and creativity.

These historical studies investigate the consequences of thinking interference freely. Rhetoric against confusion and mixture drives critics of distributed models of associative memory from the Cambridge Platonists to Jerry Fodor. It springs not only from technical concerns about how such models perform, but also from assumptions about just how confused human memory really is. Should extensive blending effects be built into our model of memory, or should order and independence among traces be taken as the natural state or competence to be explained, from which performance deviates? What features of human cognition, exactly, are defended in attacks on alleged chaos?

1.3 Historical cognitive science
Philosophical amnesia and the uses of history
D. G. C. Macnabb laconically comments (1962: 360) that ‘The unsatisfactory nature of Hume’s account of memory is noticed by nearly all his commentators. It is a fault however which he shares with nearly all other philosophers.’ Aaron (1955: 136) likewise laments Locke’s ‘slight and superficial’ treatment of memory. One does not need to think Locke or Hume got everything right to question the modern hostility to neurophilosophy which such historical judgements typify. Most early modern philosophers accepted specific accounts of the physical processes underlying and constraining cognition: modern analytic philosophers, in contrast, preferred to have no theory of memory than to rely on neurospeculations. The first full English translation of Descartes’ L’Homme, which includes his weird philosophy of the body and
the elements of his distributed model of memory, was not published until 1972, when its translator commented that Descartes' 'mechanism for explaining association is not very convincing' (H 90, n. 138). In turn, the many physiological references in Locke's Essay just disappear from some modern editions: A. D. Woozley's often reprinted edition (1964/1977), for example, omits without acknowledgement Locke's reference to tension between his account of personal identity and the animal spirit psychophysics of memory (chapter 7 below). Philosophers in both 'analytic' and 'continental' traditions have long suffered historical amnesia in suppressing and ignoring neurophilosophy. A more nuanced history of early modern attempts to deal with uncertain territory between memory and body needs to be reclaimed.

The damaging anti-naturalist desire for principled divisions between philosophy and science (Kitcher 1992: 53–9) left little room to acknowledge neurological connections in the history of philosophy. Old references to cognitive or brain processes were too easily explained away as pre-Fregean confusion of the logical with the empirical, the mistaken offering of psychological or (worse) neuropsychological answers to epistemological questions. Pitching in for further bouts of Descartes-bashing, more recent combative critics of modernity latch on to the lovely historical narrative which spies 'the original sin of modern philosophy' (Rorty 1980: 60, n. 32) in the seventeenth-century 'invention' of ideas as dubious reflections or representations of the world which cut us off from a newly veiled reality.

Naturalistic history, in contrast, does not seek the roots of unfortunate modern fallacies, but traces complex development in old cognitive theories before returning to parallel contemporary problems (Meyering 1989; Hatfield 1990). Neither naturalism nor reductionism necessarily derives from philistine carelessness about the complexity and sophistication of human history and culture. Stephen Straker suggests a negative answer to his own question: 'The problem should take the form: “could a true and coherent history of theories of perception be written from the point of view of the triumph of the reductionistvision?”' (Straker 1985: 256, n. 30).

5 The 1977 paperback edition went through ten impressions by 1981. Woozley's own discussion of memory occurs, significantly, in a widely read textbook called Theory of Knowledge (1949). For those who saw theory of memory as but a branch of epistemology, early modern neurophilosophy of memory made no sense: Macnabb (1962: 106, n. 15) puzzles over Hume's reference to the animal spirits theory of memory at Treatise I.ii.5, wondering 'why Hume thought that an excursion into physiology was more necessary in this context'.

6 Straker thinks that 'only historical studies can “rescue” philosophy from present infelicities' (1985: 272). He complains, for example, that Paul Churchland's proposals for revolution in our self-conception reveal only a failure 'to appreciate the depth and richness of the question' of the historical dimensions of subjectivity and perceptual theory (1985: 256, n. 30).
But detailed history (of a sort perhaps more familiar in the microstudies of historians of science, or in new historicist literary criticism, than in history of philosophy) can attend to contradiction, to internal fragmentation in our traditions, to the arbitrary confluences of factors driving conceptual change, and must itself be thoroughly naturalistic. The model of explanation with which the reductionist flirts is explanation which reduces, or explains in terms of something else, without explaining away. Temperature does not cease to exist just because it is reduced to mean kinetic energy of molecules. Alleged dichotomies between reductionism and the crowning glories of our species' achievements fail to justify either humanist resistance to reduction or blind scientific hostility to culture. It is precisely because history (especially the history of aberrant bodily fluids and wriggling spirits) is so resistant to reduction that it is less important strategically to beware naturalism than to deflate claims that mind and self run along autonomous from nature. Reductionism and general commitment to the 'unity of science' can coexist with acknowledgement of and pleasure in the disunity of historical and physical phenomena.

This is not of course to deny the existence of crude atheoretical technologism with no ear for history. The sensible rejection of scientism need not force us inevitably to use history against naturalism. Those who resist the encroach of science on mind, fearing that it will swamp historical awareness, are attacking a 1950s ghoul, a dreich behaviourism without laughter.

7 I use the term 'neurophilosophy' to catch common features of the two models of memory I discuss, animal spirits and connectionism, one forgotten and one fashionable. It is neutral, in my usage, on metaphysics: Descartes and Hartley gave physiological accounts of memory, but thought that some mental events were non-physical. The key is belief that 'lower-level' descriptions and processes constrain, affect, and permeate the description and functioning of psychological capacities. Although I am happy to talk about reduction (because successful reduction precisely rules out elimination), those whose metaphysics turns on supervenience-without-reduction can think of supervenience: the point is that, rather than inevitably implying eliminativism, neurophilosophy is compatible with many different views of relations between levels of explanation (Sutton 1995). Distributed models can be and often are pitched at some level of abstraction from the neural: thus weird neurological detail in Descartes or Hartley can be treated as 'brain-style' modelling, with distributed traces being neither purely psychological nor purely neurobiological.

8 In an instructive exchange with Patricia Churchland, Keith Campbell (1986) argued against the requirements of reducibility which strong naturalism imposes on particular discourses, because such requirements threaten to force the abandonment of history, literature, moral talk, and social thought: But one does not have to use Campbell's confidence in the 'human truths' of 'humanistic works of (chiefly) literature' to isolate human achievement and cognition from neurophilosophy: Shakespearean, Joycean, or Humean psychologies need not oppose sciences which wonder about strange places in the brain. Equally, Churchland's rejection of the need to understand motive, emotion, and the passions when studying brains and internal geometries (1986b: 271, n. 1) unnecessarily resists genuine pluralism in psychology, which would require history, culture, and literature too.
INTRODUCTION: TRACES, BRAINS, AND HISTORY

Metaphors and the historiography of theories of memory
Spatial descriptions of the organisation of memory, reliant on current external recording technology, defend against cultural and psychological fear about loss of control. Talk of memory as rooms, palaces, or purses, as a bottle or a dictionary, as tape recorder or junk box (Roediger 1980: 233) incorporates into body and mind ways of keeping items safe, retaining control over fluid memories. Notions about selves in relation to natural and social worlds are already implicated in theorising about memory. Ancient and medieval arts of local or place memory, for example, supplemented weak natural memory with diverse techniques for rigidly fixing items in secure artificial locations, random access systems of images for the executive self to extract at its pleasure (Sutton 1997). Psychological language assimilates descriptions of body and external world, and the process artificially sediments cultural residues in descriptions and then experiences of memory (Jaynes 1976; Scarry 1988; Johnson 1991; Richards 1992: 54–66, 104–34; Derrida 1996: 15). Quite different thoughts and actions are opened up with metaphors of memory motions and dynamic traces than with metaphors of secure independent traces located firmly in a memory bank.

It is easy for us to enjoy picaresque old accounts of animal spirits roaming the ‘cranking turnings and windings’, the ‘folds and lappets’ of the brain (Thomas Willis, in Brown 1977: 49), less so to pick up high-level assumptions already implicit in our own sciences. The ‘hardest’, least society-driven neurobiology still requires theoretical input from other levels of explanation at which metaphor and culture have a purchase. Donald Hebb believed that those who complain at the use of physiology in psychology still possess ‘an infantile neurology in [their] unconscious’ (in Colville-Stewart 1975: 410). Neuroscientists need to find out what to ask, what to look for, what to do with their data. 9 Brain research into memory, concludes Sandra Colville-Stewart in her vast survey of memory metaphors and models (1975: 417) ‘is not so much trying to find the answer to a problem, as to discover what the problem is’.

History encourages distrust of judgements which praise certain ‘theories’ of memory for their sober freedom from metaphor, denigrating the overly metaphorical nature of others. The point is not just that metaphor is generative rather than inevitably obstructive to the pursuit of truth, or that psychology in particular must necessarily adopt and adapt familiar terms (Wilkes 1975, 1988a: 198–229, 1990; Kearns 1987; Hoffman, Cochran, and Nead 1990): it is that memory is a domain in which there are peculiar problems in drawing even provisional lines between metaphorical and literal descriptions (Colville-Stewart 1975: 415). The specific difficulty in distributed models of internally

9 ‘Neurophysiology contains as an essential component a certain abstract level of description of the functional organization of the nervous system’ (Enc 1983: 298).
distinguishing memory and imagination is part cause of our general difficulty in separating the literal and the figural.

Strange continuity in metaphor and model from ancient wax tablets and aviaries is often noticed: ‘most current models of memory have been discussed under different names in earlier periods’ (Berrios 1990: 198; compare 1996: 208). Does this erode confidence in contemporary science, or reveal the pointlessness of studying historical theories? Some lament science’s lack of progress: David Krell (1990: 5, xi), describing ‘the staying-power of the ancient model for memory’, hopes to expose ‘the failure of neurophysiological research to render plausible accounts of long-term memory’.10 Others buttress current science, displacing genuine historical difference: dead scholars who had to rely merely on ‘natural observation and intuition’ are applauded for successfully identifying, ‘without the use of experiment . . . the same topics’ as those studied by psychologists after Ebbinghaus (Herrmann and Chaffin 1988: 1, 3). Psychology’s ‘old past’ disappears: one psychologist, overtly aiming at a ‘critical survey of the various hypotheses [about memory traces] proposed during the past 2500 years’, devotes only six pages to theories before 1900, justifying the brevity in his treatment of ‘the days before physiology was established on a firm scientific basis’ by dismissing most prior hypotheses as ‘at worst idle speculations and at best lucky guesses’ (Gomulicki 1953: vii). It should, surely, be possible to avoid both the use of history merely to undermine science, and the use of science as battering-ram to rubbish history.

Early modern theories of memory suffer terribly in historical surveys. There are clear philosophical analyses of ancient views (Sorabji 1972; Lang 1980; Krell 1990: ch. 1; Annas 1992), wonderful detailed histories of the ‘pre-modern’ technologies of local memory (Yates 1966; Carruthers 1990; Bolzoni 1991), of medieval memory practices (Fentress and Wickham 1992; Geary 1994) and theories of voluntary reminiscence (Coleman 1992). But then there is a gulf: Wyschograd’s (1970) account of ‘memory in the history of philosophy’ moves straight from Augustine to Bergson, Herrmann and Chaffin’s anthology (1988) jumps from Bacon across Hume and Reid to Kant, and the attention of good historians of psychology naturally focuses (often after brief ‘background’ sections on Aristotle and association) on memory research in

---

10 ‘To my own astonishment, I have found that modern scientific accounts of the earlier theories of thinking and remembering are not only oversimplified and inaccurate . . . the modern theories . . . despite their experimental, scientific, jargon, often are rather more unsophisticated than some medieval ones’ (Coleman 1992: xv, see also pp. 600–14). Coleman’s rejection of modern psychology of memory is taken further by the anthropologist Maurice Bloch, who complains that psychologists’ overt attempts to acknowledge social aspects of memory are thwarted by their use of a ‘much too simple notion of a person’ and their ‘failure to grasp the full complexity of the engagement of the mind in culture and history’ (Bloch 1996: 229, 216).
the last two centuries (Murray 1976; Marshall and Fryer 1978; Schacter 1982; Morris 1994).

Recently, late nineteenth-century French medical and philosophical theories of memory have been richly contextualised (Roth 1989, 1991, 1992; Matsuda 1993; Terdiman 1993; Hacking 1995). I seek to do the same for the period between Descartes and Coleridge. In doing so, I show that Ian Hacking's claim (1995: 203) that there was no depth knowledge sought about memory and self, 'no systematic attempt to discover facts about memory', until the 1870s is too strong. In particular, there is a longer background to more dynamic views of memory than is usually acknowledged: it is not the case that 'the attempt to conceive of memory in active terms is a relatively recent philosophical project' (Melion and Kuchler 1991: 3). Historians think of non-linear transformations more easily in the history of chemistry, where affinities, for instance, were 'forces “of a different order” than mechanical forces of impact', since the forces of affinity 'inhere in matter but are not absolute quantities; the strength of the affinity binding two substances together may change with the advent of a third substance' (Terrall 1996: 224). But precisely this kind of retroactive interference in prior associative bonds was at stake in animal spirits theory. Theories of memory in the early modern period and the late twentieth century may be more neurological in orientation than those of intervening periods (compare Rousseau 1969/1991: 4): but both cases confirm that caring for natural and bodily constraints is not condemning memory to a passive store.

Marshall and Fryer (1978: 21) asked whether, among the 'small number of metaphors' in the history of theories of memory, there is anything new or promising in recent talk, not of stores, libraries, or records, but of optical holographs as models for memory, models which allow for interference between memories and for redundancy of coding. How do connectionist models relate to traditional accounts of storage? In the early days of new connectionism, Rumelhart and Norman (1981: 2) wondered if distributed models of memory might ‘offer an alternative to the “spatial” metaphor of memory storage and retrieval’. In using history actively to address these issues, offering retrospective analyses of old theories on the basis of new alternatives (compare Patton 1994: xi), I hope to find more subtle views in the past than are perceptible with the eyes of a straight historian of memory.

In a study of nineteenth-century views of 'the double brain', Anne Harrington acknowledges that her historical perspective could be immediately relevant to contemporary studies of hemispheric specialisation. But she retreats, noting that 'the historian must make every human effort to discipline his or her culturally colored subjectivity and take the historical evidence on its own terms' (1987: 5). I refuse this sensible historian's caution, and must flirt throughout with the twin dangers of nostalgia and present-centredness.
1.4 Self, body, memory, control

As I show in part II, early modern philosophers, aware of but disturbed by the dynamics of memory, held strange beliefs about the relation between self and body. Like contemporary researchers on intentional forgetting but with extra attention to memory’s bodily bases, they thought that training in the search after truth, the inculcation of cognitive discipline, can teach us to influence or directly act upon our own brains and spirits, to achieve true virtue by the exercise of moral dominion over our own bodily fluids. After the Fall, we cannot simply ‘erase the brain’s images’ and ‘instantaneously arrest the disturbance in the brain’s fibers and the agitation of its spirits merely by considering [our] duty’: our efforts ‘to combat licentiousness’ must now be indirect, through a neurological ethics (Malebranche, LOv. 4: 360, 357). Animal spirits tended to violate the moral agent’s decency, and philosophers aware of the threat could not afford to ignore the microphysics of human nature. Moral physiologists attended intensely to methods and stratagems for gaining control over their own body processes. Memory may seem isolated from mechanisms of social control (Richards 1992: 102): but ways of remembering (as well as ways of thinking about memory) efficiently internalise norms in the body (Connerton 1989: 102; Strathern 1996: 28–36).12

After Wittgenstein, many philosophers mock these beliefs, just as they deny that action requires us to will our own body to move. They deny the intermediate steps of representation between intention and action, and blame the passive mechanistic conceptions of the body invented in seventeenth-century philosophy for such errors. But the body theorised by early modern neurophilosophers was never just an inert house for a ghostly soul. The body’s fluids and spirits, and the traces it conceals, were always active, always escaping notice, always exceeding the domain of the will, always giving shape and flavour to the soul’s plans.

It is too easy simply to say that there is no problem about the relation of self to body, or of self to memory. Dropping the dualists’ ghostly soul acting behind and through dead flesh does not deliver up a final easy naturalistic view of action. All the difficult issues remain. We do seek to change relations between self and mental contents, and between self and body. Even if we are, in part, our

11 Compare Kihlstrom and Barnhardt (1993) on the ‘prospects for strategic control of memory’. It is not that the idea of the self-regulation of memory is incoherent: remembering is, as they point out (1993: 114) a skilled activity, powerfully affected by early narrative training (Fivush and Reese 1992), and is constantly modified. But control of those modifications is desperately difficult.

12 Initial anthropological interest in connectionist models of memory is due to a desire to understand individual differences in response to cultural norms, without seeing social systems as monolithic hidden forces: socialisation is not the loading of a set of instructions, but the gradual build-up of specific, nuanced associative links into more stable representational structures. See, for example, Lawson 1993: 202–5; D’Andrade 1995: 136–49.
memories, and do not merely possess them, there are still dislocations between our parts, fragmentations, requiring philosophical and practical attention. So to think of memory as not archival but reconstructive is not to think of simple subjective ‘spontaneous, alive and internal experience’: instead, control in or of memory is always problematic, as Derrida suggests in calling the psychic archive ‘a prosthesis of the inside’ (Derrida 1996: 11, 19).

So philosophical theories of personal identity need not rigorously divide psychological from physical criteria for sameness of personhood over time. Memory requires the body’s role in psychological continuity too: both historically, as early modern philosophers were less forgetful of physiology than Nietzsche thought, and theoretically, in attending to empirical constraints on the extent and nature of executive control. It is hard both to take psycho-physiology seriously and to see rational inference as the basic characteristic of human cognition. Something has to give. Molly Bloom, not Sherlock Holmes, is the fictional figurehead for the neurophilosopher (chapter 11 below; contrast Fodor 1985b/1991: 39–41, 1987: 18).

I have no positive account of the true link between memory and personal identity. Marya Schechtman (1994) argues persuasively that facts about autobiographical memory disrupt philosophers’ accounts of continuity of self, accounts commonly used to buttress theories of agency and moral responsibility. She argues that ‘the immense complexity of the relation “memory of”’ (1994: 9) challenges the Lockean assumption that all we, courts, or gods have to do in deciding problem cases about sameness of personhood is to check up on simple connections between two well-defined (past and present) moments of consciousness. Schechtman finds in empirical work on reconstructive personal memory more realistic pictures of the roles of narrative continuity in personal identity, acknowledging the ways we construct and change the past to make it ‘more smooth and comprehensible’.

It is precisely insofar as our memories smooth over the boundaries between the different moments in our lives, interpreting and reinterpreting individual events and experiences in the context of the whole, that we are able to produce a coherent life history. It is by summarising, condensing, and conflating the different temporal portions of our lives in memory that we are able to see them as parts of an integrated whole, and this integration blurs the distinction between different moments of our lives. (Schechtman 1994: 13)

Revised notions of self, which do not impose unity and continuity in advance, are obviously desirable. But I do not pursue this project here. I am less concerned with new decentred conceptions of subjectivity, with deciding in the abstract whether ‘the subject’ should be discarded or merely fragmented and dissolved, than with firming up historical and conceptual connections between certain accounts of memory and certain pictures of control. Only if
memories are local items waiting to be scanned and dealt with is there need for a strong conception of an active, evaluating, transcendent self.

Some of these remarks may seem irrationalist, and they are certainly intended to undermine logicist overconfidence in the 'natural order' of cognition and memory. But there are still, I repeat, constraints on confusion. Not everything is equally interconnected, in memory or in the self (compare Cherniak 1986: 49–71 on memory holism and rationality). It is never a matter just of doubting at will, rather of acknowledging existing cracks in cultural consensus, reclaiming forgotten alternatives, buttressing independently motivated revisions with new ideas in cognition. The ways that representational spaces in memory get partitioned are not chosen by a self which transcends them, but neither are they entirely chance, for they are the sediment of a particular past in a specific brain and body. Alien memories, like feelings, can also be our own, and strangeness can lie inside (Lingis 1994b: 99–103). This is why memory has its sadness: because remembering is reconstructive, it is also destructive. Like anatomy, memory mangles and transforms its materials, tending to obliterate as well as construct. It is not only in repression and the organised forgetting imposed by oppressive regimes that memory is continuous with violence: there is also, as Francis Barker argues (1994: 85–6), a 'more rarely noted violence of recall'. Science does not inevitably neglect such cares.

I should say, finally, which of the many forms of memory distinguished by some psychologists I am concerned with. Superposition and distributed representation are particular mechanisms of persistence which need not apply across every domain of memory, with the extent of their application to be decided only by empirical research in neuropsychology across various systems and subsystems: the 'general theoretical proposals and ideas about the nature of memory' put forward by the authors I discuss can be suitably restricted when the range of superposition is decided (Schacter and Tulving 1994: 1). My interest is primarily in the subcategory of episodic memory which includes autobiographical remembering of specific episodes from a personal past (Larsen 1992). But I do not restrict attention to those personal recollective memories which are 'a "reliving" of the individual's phenomenal experience during that earlier moment' (Brewer 1996: 60), since many important effects of the personal past operate outside immediate awareness: thus these theories also cover implicit memory, the unconscious influence of past experience when the past, so to speak, leaks into, or contaminates, behaviour and ongoing cognition (Schacter 1995: 19). Further, I include passive remembering, in which events come to mind unrequested (Spence 1988). Autobiographical memories are not necessarily marked as such, nor do they come neatly dated (Larsen, Thompson, and Hansen 1996): there is room for the idea that some memories become autobiographical in use and improvisation, on the basis of both cultural and individual narrative norms (Barclay and Smith 1992: 75; Barclay 1996).
Appendix: memory and connectionism

For readers with little previous exposure to connectionism, I append a tiny simplified sketch (good introductions to connectionism and memory include McClelland and Rumelhart 1986; Smolensky 1988; Clark 1989: ch. 5; Churchland and Sejnowski 1989, 1992: chs. 3–4; Bechtel and Abrahamsen 1991: chs. 2–3; Rumelhart 1992; O'Brien 1993; Collins and Hay 1994; McClelland 1995). I do not distinguish among its many varieties, but should note that some recognisably connectionist models do not use distributed representation: I ignore them here.

Minimal distributed connectionist memory networks involve the following ingredients. First, processing units have a continuous range of activation values, for instance between $-1$ and $1$. Second, a (physical) pattern of connectivity between these units determines which units send output and receive input from which others. The pattern of connectivity can be unidirectional (feedforward nets) or can allow for feedback, via mutual input/output connections between units or through introducing extra context layers of units. Third, weights on each individual connection play a part in determining the transformation of the net output from one unit into input to another.

These connection weights between units change in the course of processing, usually after a stable activity pattern is achieved, according to one of a variety of learning algorithms. A unit's activation value is computed from the net input values communicated to the unit from the other units to which it is connected. This net input can either simply become the new activation value, or, more interestingly, can combine with the current activation value of the unit, the combination perhaps including a bias or threshold function, allowing non-linearity or 'squashing functions' (Churchland and Sejnowski 1992: 62–5, 109–12), by which small changes in input can cause large-scale shifts in activity. The gross activity pattern of a simple network (the whole set of individual activation values) is determined by the combination of the existing pattern of connectivity, the weights on these individual connections, the present activation values of the units, and the present inputs to the system.

Simple computation in such a network is the processing of continuous inputs so that the pattern of connectivity weights changes to produce a global pattern of activation which satisfies as many of the mutual constraints between units as possible. The network 'relaxes' into a stable state, settling into a solution through iterative operations performed in parallel by units which are only locally

---

13 As I am not defining differences between 'classical' and connectionist models of memory, I do not mean to overdo contrasts between them. The strategy is to develop on its own terms a connectionist approach to cognition, to see what such a thing might possibly be (van Gelder 1990c). Apart from the treatment of Fodor in part III below, I contribute to direct criticism of 'logicist cognitive science' merely by showing that connectionist-style theory was once common, submerged only by contingent historical changes.
connected to others and are under no global executive control. ‘Remembering’ is, like other operations, ‘a process for producing an output’ (Wiles 1994: 80).

How does memory occur in these mechanisms? When inputs are continually presented, patterns of activation come and go. But the shifting patterns leave traces behind them (McClelland 1995: 69–70):

Forming a memory trace for something – say, an episode or event – begins with the construction of a pattern of activity over the processing units, with the experience itself strongly influencing the pattern. But the existing connections among the units will also influence the pattern constructed, thereby introducing the possibility of additions, omissions, and distortions. Storage of a trace of the episode or event then occurs through the modification of the strengths of the connections among the units.

Traces are in the interrelations between units rather than fixed to individual units, and ongoing processing takes place in the very same (parts of a) system as ‘storage’.

New input is not just accumulated, preserved in an unprocessed form, but is filtered through existing representational space in which many traces are compacted (Touretzky and Pomerleau 1989; Elman 1993: 79; Wiles and Ollila 1993). So ‘search’ through memory cannot be a systematic scanning from a transcendent command viewpoint of many independent memory locations. Memories do not need specific addresses in order to be successfully retrieved. The indexing schemes of the traditional spatial metaphor break down if there is any error in the retrieval cue. But in connectionist models, the same activation pattern can be achieved in response to a number of different partial input descriptions, even if noisy or degraded: if some part of a distributed trace becomes active, the whole trace will tend to activate (Anderson and Hinton 1981; McClelland, Rumelhart, and Hinton 1986: 25–9). Without having to fix external context tags to every trace, context effects are built into the system of co-occurring units.14

Distributed models thus naturally exhibit causal holism. The representations or traces ‘stored’ in the same physical system are all automatically causally implicated in all the system’s behaviour (O’Brien 1993). This is just because they are all in the same space and any reconstruction of one activity pattern is affected by everything else that has happened in that space. No network acts in freedom from its history. In part I, I show how this causal holism operated in the brains and bodies described by animal spirits theory.

14 It is worth noting, in addition, the importance of the absences in connectionist models. The lack of central executives, explicit rules, explicit prototypes, discrete memory addresses, and external context tags has given these models at least some appeal even to philosophers who have been scathing about the claims of classical artificial intelligence to model human cognition (Dreyfus and Dreyfus 1988).