1

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

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 pretence 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 neuro-philosophical 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 *mech-*

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

of psychology (Churchland 1993: 218–19, 1995: 286–94).[1] 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

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 identity, 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),

> once a given pattern has been processed and the network has been updated, the data disappear. Their effect is immediate and results in a modification of the knowledge state of the network. The data persist only implicitly by virtue 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 be context-insensitive (Serres 1982; Kirsh 1990: 342–60; Clark 1993: 121–7).

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 deliverances 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 disputable. See O'Brien 1991, and on reconstruction McCauley 1988.

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