6
Local and distributed representations

The distinction

What use is all this history? My accounts of writers who did not have available the explicit categories of local and distributed representation should fulfil two demands. Already some otherwise invisible disputes and controversies in historical theories of memory have been brought to the surface. But further, the metaphysical and scientific utility of the local/distributed distinction itself should be illuminated by these old ideas. In this chapter, then, the focus shifts briefly from historical responses to the animal spirits model of memory to conceptual evaluation of its implications. I introduce the key distinction within distributed models between explicit and implicit representation, and conclude with a speculative mapping of the analogy between animal spirits and connectionist versions of distributed representation.

It is not obvious that dividing distributed from local models of memory is, even now, a sensible thing to do. Even if there is a genuine distinction, it cross-classifies ideological divisions between connectionist and classical cognitive science: not all connectionist models are distributed. More importantly, the distinction seems to be only perspectival, describing the ends of a spectrum of possible models rather than well-defined alternatives (Clark 1989: 95). As Churchland and Sejnowski remark, 'On the face of it, the difference between local and distributed representations . . . falls rather short of spine-tingling, barn-burning stuff' (1992: 163). Like them, however, I think that the utility of the distinction can be shown (see also Lloyd 1989: 102–16). I test it in historical practice, by setting up possible applications and arguing for my own.

Hooke's view that memories are 'in themselves distinct', I suggested, is characteristic of localist approaches to memory, in which one unit codes one item. But Jamie Kassler, surveying seventeenth-century theories (1995: 129–39), takes Hooke to have a partly *non*-localised theory of memory. This reveals alternative ways of drawing the local versus distributed distinction, through alternative meanings of the term 'local'. Kassler (1995: 113–15) sees the distinction as marking the *amount* of body and brain substance involved in storage. This brings the issue close to the traditional problem of whether the soul is coextensive with and spread (distributed) throughout the body, or is localised at a particular seat (French 1969). Because Descartes has the soul interacting with the body only at the pineal gland, this scheme assigns him a *local* model of memory. There would be no *special* difference between local and

distributed memory separate from questions about the manner of psychosomatic interaction. On Kassler's account, the opposite of 'local' is not 'distributed' but 'outside the brain'.[1] Descartes' statement that memory impressions are 'mainly located in the whole substance of the brain', and his explicit denial that they are exclusively located in the pineal gland would not mitigate his localism.

I distinguish local from distributed models quite differently. I defend my distinction not only because it is more in line with other uses of the distinction in new connectionism and the cognitive sciences, but because it makes more sense of the historical theories. The distinction as I use it, relates to the nature of the storage involved *wherever* memories are stored: it is about the discreteness or independence of memories one from another. In any model, do memories remain separate, or do they naturally combine?

The question of whether the soul had a particular seat or was dispersed through the body is not, in my view, at issue in the specific case of memory. Accepting, as Descartes did, that memory can be in the hands, say, or that other parts of the body, like the heart, can *influence* remembering is not directly relevant. *Everyone* agreed that relations between brain and the rest of the body explain peculiarities or abnormal cognitive and behavioural functioning. For Descartes, as I showed in chapter 3, the nature of the animal spirits arriving from the heart is a vital factor in ongoing cognitive processes, and their nature is affected in all kinds of ways by a wide variety of bodily states and processes, such as stomach juices, the air taken in in respiration, and the states of the liver, gall bladder, and spleen (*L'Homme* AT xi.168–9, H 74–5). Nobody saw the brain as an isolated, independent unit. But the mere influence of the rest of the body on remembering is not enough to make a theory non-localist: indeed only local memories (in my sense) remain carefully bounded in their memory places.

The distinction is not to do, in turn, with the *amount* of brain substance involved in memory storage. Distributed traces are compatible with *macroscopic* localisation of function in the brain. Localisation in the latter sense does not get neurophilosophy very far: 'knowing roughly where a process takes place in the brain typically tells us very little or nothing about the . . . mechanics of the process' (Hatfield 1988a: 727). There may be extensive systems of relatively independent localised modules, devoted say to mental images or to parsing sentence structure, some or all of which still employ distributed representations. As Hinton, McClelland, and Rumelhart say (1986: 79), 'The distributed representations occur within these localised modules . . . the representations . . . are local at a global scale but global at a local scale.' It is globality at the local

1 Hobbes is seen as a non-localist because he takes remembrance to be extended through the body and recollection to result, partly, from motions transferred from the heart by the *pia mater* to the brain.

scale which makes them distributed. Again, there is a real issue here about whether memory is separate from any 'executive', or whether storage is separate from processing. But this is not the local versus distributed distinction, although it is connected (as in Hooke's preference for a strong division between an executive 'center' and locally stored memory ideas).

When Descartes says that the 'folds of memory' are 'mainly located in the whole substance of the brain', this is not *sufficient* to prove his model to be distributed, though it is suggestive. But there *could* be local representation occurring in every part of the brain: and this indeed is the case in Hooke's theory. What more then is required for distribution? Memories must be superpositionally stored (van Gelder 1991b: 36–45, 53–4), many in the same place. The right sense of 'local' to contrast with 'distributed' is where local means 'discrete, separated or nonoverlapping' rather than a sense of restriction in extent (van Gelder 1991b: 42).

Room in the brain

This catches what is distinctive in the Cartesian view: Descartes bypassed, and Malebranche dismissed, the problem of finding room in the brain for every memory, by suggesting that a single fold of the brain holds many traces. John Hawthorne (1989: 9) argues that the room-in-the-brain argument favours contemporary connectionism over local models, which entail that I can only entertain as many propositions as there are neurons in my head. Does Hooke's local model bypass this difficulty?

Boyle roughly calculated the number of distinct words and things an aged learned man might remember, to understand 'how in so narrow a compass, as part of a Human Brain there should be so many distinct Cells or Impressions as are requisite...' (1772/1965: VI: 742). The problem arises with force only on the assumption that impressions *must* be distinct. When Hooke outlined his view of memory, he worked out in more detail how many ideas one person might store in a whole life. He started working on one idea per 'moment', but revised this frightening calculation down markedly to end with a figure of 1,826,200 ideas over 50 years (compare Rose 1993: 90–1 on modern versions of this peculiar quest for the total number of separate items an individual memory might retain). But even if we could 'remember 100 Millions [of distinct things], and consequently must have as many distinct Ideas', Hooke thinks that this number may 'actually be contained within the Sphere of Activity of the Soul acting in the Center' (LL 7.4: 143), and simply concludes that 'we shall not need to fear any Impossibility to find out room in the Brain where this Sphere may be placed, and yet find room enough for all other Uses' (LL 7.4: 144).

MacIntosh and Kassler explain Hooke's optimism here by referring to the previous lecture 6 of the *Lectures of Light*, in which, impressed by his microscopic observations of innumerable infinitesimally small organisms, Hooke suggests

that there might be no limit to the smallness of a piece of matter which could yet contain enormous numbers of distinct things:

> there are in every *sensible* Point of Matter a sufficient number of distinct Particles to convey every one of those Motions distinct, without interfering one with another: For as there may be Millions of Motions communicated to a *sensible* Point, so there may be as many Millions of distinct Particles to receive each of them distinctly. (LL 6.5: 134; compare MacIntosh 1983: 347–8)

Interference in local models

Hooke's later statement in the lecture on memory that the 'material and bulky' ideas of memory must be 'in themselves distinct . . . actually different and separate one from another' echoes this passage. Hooke rules out of his theory of memory the interference which Descartes and Malebranche expected as a consequence, albeit a potentially dangerous one, of their distributed model. Hooke does not even canvass the possibility of non-local representation (in my sense): memory ideas just cannot ever be 'two . . . in the same space' (LL 7.4: 142). All Hooke's memories are always explicit, lurking in the coils of the brain. The only difference possible in their state is when the soul in its circular course from 'the Center' finds and uses one of them in its processing or reasoning before returning it unchanged to its own memory spiral.

But a further look at Hooke's account may suggest that my analysis is mistaken. Hooke, in fact, *does* think that his model allows for interference. Kassler, using her own local versus distributed distinction, accordingly credits him with a successful account of non-catastrophic interference which avoids the terrible disorder and confusion which so concerned Glanvill and More (Kassler 1995: 135–6). But Hooke's claim is misleading, for his is not genuine interference. He needs to account for various phenomena of which one with such a sense of the fragility of memory was well aware: forgetting, for instance, occurs when material ideas decay in the very furthest orbs of the memory coils. Ideas are material and so subject to change: those which, 'shifting and changing place in the Repository', get 'closer and closer stuffed and crouded together' can 'be in time alter'd, and sometimes quite lost' (LL 7.4: 144). This alteration of memory ideas may look like interference, and might be used to account for the blending and generalising tendencies of human remembering. But Hooke's continuing exposition shows that his localism carries more weight than the phenomena of interference.

Other memory ideas, says Hooke, interpose between the memory sought and the centre which seeks, as a physical 'Impediment to this Radiation of the Soul' in remembering. He compares 'the manner as the Earth interposing between the Moon and the Sun, hinders the Sun from radiation upon the Moon' (LL 7.4: 144). This echoes Hobbes' notorious account of 'decaying sense': 'this "decaying" must actually be seen as an ellipsing' or occluding, since bodies

keep moving until hindered (Pye 1988: 289–90; compare Hobbes 1651/1968: 88–9, 657–8). There is no fusion of memory motions here, only a particular spatial juxtaposition of bodies, which, from the point of view of an assumed inner viewer, obscures one of them.[2] Executive access to a bit of memory information is blocked by the interposition of a new or another memory.

Hooke denied that his kind of 'interference' would be catastrophic on the grounds that the parts of the brain involved in memory are so incomprehensibly small that there is room in the brain for all of them to be separately stored. He accepts that sometimes ideas which have not 'kept the same Order in which they were made' can 'intrude and thrust in themselves between . . . so as often to interrupt and break the Chain or Order of Insertion' (LL 7.4: 144). But even such 'reaction and repercussion', such violation of order, is done in an orderly manner: the ideas never lose their own identity, but remain always, even in intrusive insertion, 'in themselves distinct'. What he calls 'the Interposition of other Ideas between the Center and the Idea sought' (LL 7.4: 144) still operates by local representation, and cannot provide for any genuine mingling of memories to the extent that they might lose their original identity. Hooke's memories remain independent of each other, with their own distinct constitutions and motions.

Implicit and explicit representation

Thinking of memory as a motion in the nimble spirits rather than a body meant that remembered items could not all be continually represented in explicit formation. At any time most memory patterns will not, in one sense, actually be present: all that is there are physical dispositions for them to be recreated, just as the linen cloth which has had many patterns of holes traced through it has dispositions for the easy reopening of particular patterns. The ontological status of the memory patterns which are not actually present at a time may seem unclear: and yet the dispositions are real, for brains or cloths with different histories will not recreate the same patterns.

One way to spell out the strange metaphysics of the distributed memory trace is to invoke a distinction between *explicit* and *implicit* representations. Though developed for new connectionist models (O'Brien 1993; compare Hatfield 1991: 95–6; and Churchland and Sejnowski 1992: 165–70), this distinction is also fruitful when applied in historical cases: in chapter 7 I will show that Locke

2 Digby, likewise, wants to incorporate an account of forgetting by fusion and loss of memories (rather than their mere obstruction from executive view) into his localist model. But fusion is impossible if memories are bodies: the closest they can come is 'coupling'. Despite rhetoric of mouldering and defacing ideas in memory, Digby is consistent in not allowing genuine interference: bodies only take on 'a maimed and confused shape in the memory' when shocked by collisions with other bodies (TT 33: 287).

clearly understood it. Local models have no implicit representations, and must get by with enduring static explicit representations. In contrast, explicit representations in distributed models are passing patterns of activity, evoked across neuronal units or animal spirit motions by the combination of previous activity pattern, patterns of connectivity and connection weights, and present input. Any one system can only be in one state at a time, and cannot simultaneously display multiple patterns of activity.[3]

So when I describe philosophers thinking of memories as motions, I mean that they take the explicit tokening of a pattern of activity to be, or to be correlated with, the occurrent remembering. Remembering the analogy with light (chapter 4), however, we might prefer to see memories as tendencies to motion. It is in this latter sense that there can be many memories overlapping in the same place, as implicit representations which can all potentially be rendered explicit or actualised.

Where explicit representations are transient, *implicit* representations endure. They are dispositions which allow for or ground the recreation of the explicit patterns, the changes in the connection weights or brain pores without which such reconstruction would be impossible (McClelland and Rumelhart 1986). Of course there is only one set of values in the weights at any one time: so to be precise some speak of many representations in one 'representing' (van Gelder 1991b). Indeed, on some views, only implicit representations are strictly distributed: explicit representations, some argue, are 'functionally discrete' just because there is only one at a time in a system (O'Brien 1991, against Ramsey, Stich, and Garon 1991). I do not pursue this debate because the difficult consequences of superposition arise even when attention is restricted to implicit traces, multiply coexisting in the same space at the same time.

What is at one time explicit may in future processing become implicit. Information represented implicitly is tacit, 'in the weights', only potentially active. This is the way in which, most of the time, we 'store' our telephone numbers: but on request, it comes to be explicitly represented. In turn, what I have been calling 'reconstruction' is the change in state from implicit to explicit, the actualising of a disposition. As the example of the phone number shows, confusion does not inevitably result: I usually remember where in the car park I will find my car without confusing today's location with every

3 This does not rule out the 'explicit' remembering of many things at once, for one organism or machine may have many systems. Further, there may be inconceivable rapidity of change in the patterns of activity within one system as one explicit trace shades or shifts into another or between many: temporal dynamics account for many of the phenomenological effects of association. However, it is not easy to spell out just what 'explicitness' requires here, or in particular how it relates to consciousness. I avoid this issue: for speculation see O'Brien and Opie forthcoming. My interest in this book is primarily in implicit traces.

previous day. The question is whether these examples tell us much about more complex autobiographical remembering.

Burdens of explanation

So distributed models must allow the occasional retrieval of distinct memories: local models must allow occasional fusion between, or generalisation across, memories. Local models take the distinctness of the 'figures', 'motions and constitutions' of memory traces as primitive, whereas a distributed model takes as primitive the particular way superpositional storage occurs in the physical substrate. Can distributed representations effectively avoid catastrophic interference so as to approximate the faithful recall which we sometimes achieve? Can local representations gerrymander blending and context effects by adding extra twists or mechanisms of the kind we found in Hooke?

Our intuitions about local representation fit ordinary digital computers. When each memory is discretely stored, unchanged in its location until recalled by the central executive, information is faithfully reproduced, unaffected by intermediate experience. The memory brought out of storage into working memory or a temporary buffer is a duplicate of that which was originally encoded.[4]

Disputes between the models can also address the nature of the alleged phenomena of remembering which they seek to explain. If humans almost *never* claimed to be remembering when in fact confabulating or spuriously reconstructing, then the local model of memory would look much more attractive. The human mind/brain would be much more like other storage mechanisms, external means of recording and keeping information safely. It is vital, *for us*, that our word processors store discrete, independent files in identical form overnight: any interference, blending, or mutual contextual changes of different files stored on the same disk would be disastrous. Much writing on these subjects, whether it acknowledges this or not, has been trying to characterise the explananda suitably and convince others that some phenomena are more theoretically important than others. Should cognitive theory *impose* order, or allow order to emerge?

Animal spirits and neural nets

To sum up with irresponsible anachronism, I conclude by fitting animal spirits into the conceptual framework also exemplified by connectionist models

4 Advocates of distributed memory can argue that the decision to call the localist storage system of the digital computer a 'memory' was little more than an unfortunate bad pun. Babbage called his recording device a 'store', and the storage system of the American wartime computing project was not called a 'memory' until John von Neumann introduced the term as part of an explicit, forced analogy between computers and human brains (Bolles 1988: 166–8).

(O'Brien 1993). Distributed representation operates at a level of abstraction from specific neural matter.

Abstract feature	*Parallel distributed processing*	*Spirits*
ARCHITECTURE		
Processing units	'Neurons'	Brain pores
Activation value	Spiking frequency	——
Range of inputs	——	Spirit flow
Output	——	Spirit flow
Networks of units	Neural nets	Regions of brain
Pattern of connectivity	——	Structure of pores
Mechanism for plasticity	Connection weights	Microstructure
(DISTRIBUTED) REPRESENTATION *(Information is encoded through plasticity)*		
Explicit representations ('traces'?)		
Transient	Patterns of activity	'Figures traced in gaps'
Extended	Vectors	(Implicit in sensory isomorph)
Implicit representations (= traces)		
Enduring	Modifiable connection weights	Altered pores
Superpositional	Single weight matrix	Many in same region
PROCESSING/COMPUTATION		
Method of computation depends on architecture		
Processing in both cases is analog: 'rules' for computation are physical causal laws; thus changes in operation are changes in the substrate		
Both models exhibit causal holism by which all superposed traces influence product of any processing		