notwithstanding its highly original insights and its numerous thought-provoking *aperçus*.

Department of Philosophy,
Leiden University,
The Netherlands.

Author's Response

By John Sutton

Historical Cognitive Science

I am lucky to strike three reviewers who extract so clearly my book's spirit as well as its substance. They all accept and act on my central methodological assumption: that detailed historical research, and consideration of difficult contemporary questions about cognition and culture can be mutually illuminating. It's gratifying to find many themes which recur in different contexts throughout *Philosophy and Memory Traces* so well articulated by my reviewers. They catch my desires to interweave discussion of cognitive theories of memory with moral questions of psychological control and self-mastery, to evoke the virtues and the pleasures of strange, baroque beliefs about fickle 'animal-spirits' coursing through the nerves and the brain, to demonstrate that mechanistic explanation (even in its blunt old Cartesian form) can acknowledge complexity, and to develop scientific conceptions of dynamic memory traces and representations which can survive uncharitable philosophical criticism. The book's insistent interdisciplinarity is just an inchoate quest to acknowledge the daunting variety of the phenomena: remembering is both natural and cultural, and is studied by narrative theorists as well as neurobiologists, by physicists as well as psychologists. By fusing the detail of a history of early modern neurophysiology with the committed, even gullible fervour of a defence of 'new connectionist' cognitive science, I wanted to pull out the carpet from all those who are happy to let 'scientific' and 'cultural' approaches to the mind run along independently. Once this general project is given space, as it is by all three reviewers, we can get down to specifics.

From Neural Nets to Animal-spirits

The book describes and defends two reconstructive theories of autobiographical memory. One is the bizarre old neuromythology of memories as dynamic patterns in fleeting 'animal-spirits', nervous fluids thought by early modern philosopher-physiologists to rummage through the pores of brain and body. The other is the connectionists' 'distributed model of memory' developed over the last twenty years. What do these two theories, one forgotten and one fashionable, have to do with each other? Are the similarities between them deep, or merely the illusory product of indulgent anachronism?

My primary energy was devoted to the seventeenth- and eighteenth-century versions of the animal-spirits theories of memory, displaying the intricate range of contexts in which these fleeting body fluids entered early modern theory and experience. The easy reach of such superficially silly old theories across natural and cultural domains, which we have come to see as distinct, might intrigue readers who want to see our own cognitive sciences reach out from their theoretical and institutional limits to tell those on the outside things they want to know. The force of my historical analogy derives, I hope, just from the juxtaposition of old and new concerns about truth in memory, about control of the personal past, and about bodily constraints on cognitive discipline. But in a more ambitious historical agenda, and in full awareness of those dangers of present-centredness which Theo Meyering mentions, I also argue that the structural similarities between the two theories of memory are not superficial. Some *historical* insight is gained by thinking of these outdated models of memory with the new connectionist perspective in mind; and some *philosophical* insight is gained on current scientific debates by examining resistances to, and perceived consequences of, relevantly similar old theories on which history affords us distance.

But what detail can I provide on what Meyering calls these "abstract but highly general" parallels between the two theories? (First let me take Catherine Wilson's point that a more extended basic explanation of the modern models of memory would help: my primer on connectionism for historians and social theorists—the audience I'm trying to convince to take cognitive science seriously—could usefully have been both longer and illustrated.) The explanatory core of the general 'distributed' model which is common to both, which generates the characteristic *plasticity* of memory in connectionism and in animal-spirits theory, is indeed "at a decidedly abstract level" (Churchland and Churchland 1996, p. 225), and does not depend on the details of a single particular physical realisation.

This is a two-factor picture by which transient 'rememberings' are generated in, and out of, an enduring system. As two of the leading exponents of the connectionist version put it:

> what is essential is the idea of fleeting high-dimensional patterns being transformed into other such patterns by virtue of their distributed interaction with an even higher-dimensional matrix of relatively stable transforming elements. The fleeting patterns constitute a creature's specific representations of important aspects of its changing environment. And the relatively stable matrix of transforming elements constitutes the creature's background knowledge of the general...features of the world (Churchland and Churchland 1996, p. 226–7).

On this abstract theoretical apparatus, when one occurrent representation is *explicitly* present (in, say, actually remembering), the system which generates it also holds many *implicit* representations, 'stored' superpositionally (for instance, in the connection weights of a trained neural network).

But what could any of this *amount to* in such a strange old theory as the one I attribute to the animal-spirits theorists? Well, the transient explicit representations are what Descartes called in *L'homme,* his weird treatise on the philosophy of the body, 'figures traced in gaps', patterns of fluid flow between brain pores in a neural tissue. Meyering asks what is equivalent to the activation values of the processing units (values which compose a vectorial pattern of activation in neural nets). My answer (perhaps not that clearly portrayed in my diagram on p. 156) is the *continuously-varying* values characterising the flow of animal-spirits: for Descartes, that is their direction, strength, and consistency. Except for some wonderfully optimistic early eighteenth-century 'Newtonian' physiologists, nobody tried to quantify these variables, and so there is of course *one* obvious sense in which the mathematical framework of modern neural net theory makes it utterly different. But there is another sense in which 'patterns' in both frameworks are realised in analogue (rather than digital) physical media, allowing an infinite range of states. Despite the vaunted 'neural fidelity' of new connectionism, it isn't tied only to our particular technologically-driven conception of neural networks: distributed representation itself exhibits what philosophers call *multirealisability,* with the same dynamics being generated in quite different physical systems. This isn't a functionalist rejection of mechanistic explanation: the particular physical system in question will have to be understood in order to follow the kinds of representational transformations which it realises. When I argue in Chapter 13 that David Hartley, who rejected animal-spirits in

favour of vibrations along nerve fibres, nonetheless also had a distributed model of memory, I'm claiming that he too was committed to *some* set of continuous physical variables (notably what he calls the 'strength and frequency' of vibrations) which serve to realise the same abstract two-factor model of transient explicit and enduring implicit representations.

Meyering also requests clarification on what the early modern equivalent might be to the modifiable connection weights that are the mechanism of plasticity or learning in neural network models. My summary reference to the 'microstructure' of the brain on p. 156 is shorthand for the whole pores-and-fibres fantasy of those who accepted what David Hume called "the Cartesian philosophy of the brain". So where the role of fleeting patterns is filled by the changing flow of animal-spirits, the role of 'relatively stable transforming elements' is filled by the fibres of the brain's 'complex net or mesh', through the pores between which the nervous fluids roam. These fibres are flexed, enlarged, constricted, bent and rearranged by the spirits over time, and can retain the flexures received in the course of this experience, in such a way that many previously-existing patterns of spirit flow can be recreated (on appropriate input), 'without', as Descartes says "requiring the presence of the objects to which they [the patterns or traces] correspond. And it is in this that *Memory* consists". There is no clear distinction here, as is made in most new connectionist models, between the basic architecture of the system (its physical layout or pattern of connectivity) and the enduring but modifiable weights on particular connections: but it's far from clear now just what the biological reality of this distinction will turn out to be. And this difference is anyway minimal compared with the striking recognisability of the pores-and-fibres picture of plasticity, as the means by which the particularity of past experience is carried in the brain.

Conceptual Change

Meyering's aim in pressing me on these points is to home in on deeper questions in the history and philosophy of science. If I've successfully provided the 'corresponding mechanisms' he requests, then what picture of theoretical continuity follows? I certainly accept the broad outline of the pragmatic account of conceptual change which he outlines, although I am still puzzled about, and working on, the problem of the reference of theoretical terms in a case like 'animal-spirits', where the entities in question don't exist, but where much of (scientific as well as social) value might yet be salvaged from the strangely dispersed discourses in which they were embedded. I've tried in passing to raise specific questions for historians of

concepts: placing Hartley's stress on the importance of the 'rate of recurrence' of vibrations in this context, for example, almost inevitably suggests a new perspective on the prehistory of the concept of frequency in neuroscience.

But it's not part of my case to suggest that there *must* be direct lines of 'historical influence' across theorists of memory straight down from, say, Descartes to the Churchlands: especially with a neglected and haphazard piece of the history of science like this, the long-view lines of causal/intellectual passage are way too messy and interrupted to pin down. No, the driving idea is a more brute realism about the various theories of distributed memory: both animal-spirits theorists and new connectionists are responding to, trying to describe and explain, the same phenomena. What's in common isn't just the abstract functional description of the two models, but also a shared conception of the explananda, a willingness to focus on errors and distortion in memory, on interference and blending effects, on failures of control and on the difficulty of keeping the personal past in order. Opponents of distributed representation, from Glanvill and Hooke to Fodor and Pinker, require 'the exactest order' to be preserved among the items independently assigned a location or address in a storage system (or along Hooke's 'coils of memory' in the brain). From their perspective, mixture and confusion in memory is merely a sign of imperfect performance, of unfortunate departures from an idealised cognitive competence. In contrast, the reconstructive nature of remembering on distributed models makes human memory look more like a compost heap (p. 244), with episodic memories arising holistically out of a conspiracy of implicit causes, more Molly Bloom than Sherlock Holmes (pp. 235–6).

Control and Time in Memory

Yet this dichotomy is obviously too rigid. Of course there are many more moderate pictures, and I argue in the book that the difference between distributed representation and local storage is better thought of as a spectrum. Catherine Wilson suggests that I overdo my rhetoric about the fragility of memory. I plead guilty, and agree that my overall theses do not require me to harp so much on memory's pervasive unreliability. While I am impressed by the reasoning of both Hartley (p. 256) and the connectionists that certain models of memory (like their own) may be supported by pointing to characteristic human patterns of error and distortion which these models predict, I don't mean to neglect the mundane success of memory. What I want to point to is not so much the *fallibility* of memory as our concern about lack of *control* over the processes of

remembering, especially in the realm of significant autobiographical memories. As Wilson rightly remarks "it is not only gaps and losses that trouble us; accurate memories too may be unwelcome, distressing, and difficult to integrate with one's present".

It's too easy to mock rationalist confidence that veridical memory can be isolated from the story-weaving mechanisms of imagination; just as interesting are the ways in which reality intrudes into fantasy or into abstract thought, when cognitive control is threatened by persistent rumination or intrusive recollection (Schacter [1999] describes recent experimental work at different levels, and Engel [1999] elegantly addresses the role of context in personal memory). Both animal-spirits and connectionist approaches to memory collapse storage and processing into the same system, so that there is no separate executive mechanism picking out and manipulating items at will before dumping them back into localised memory banks. This means that the same kinds of causal processes are in play, whether the current dynamics of the system are driven by external reality, by dreams, or by reason.

The explanatory burden for such reconstructive models therefore shifts, so that the most puzzling phenomena are not our occasional lapses and confusions, but our habitual and roughly veridical genuine access to the personal past. Why is there not inevitable catastrophic interference between overlaid traces? How, in particular, do we manage the remarkable feats of mental time travel characteristic of episodic memory, in which we are in psychological contact of some kind with very specific, datable events in our past, events around which we can spin our autobiographical narratives? This phenomenon exemplifies our capacity to represent what is absent. Incidentally, it also exemplifies the incompleteness of existing versions of 'ecological psychology', in which followers of J.J. Gibson either neglect such very long-term personal memory as some kind of 'luxury', or implausibly argue that all the information we need *is* in fact contained in the external environment rather than in mind, brain, or body. So while I am 'tempted', as Wilson puts it, by the dynamic picture of the 'embedded, embodied' mind in Gibsonian tradition, I reject in Chapter 15 the idea that Gibson has refuted the need for *any* representations and traces in memory. John Campbell (1997) argues that such autobiographical remembering, in which we are oriented to particular past times, requires a uniquely human rich and stable grasp of the linear connectedness of time in memory, and a strong conception of the spatiotemporal continuity of the self. It's a pressing question for a general connectionist philosophy of mind whether this is so, or whether our narrative grasp of datable past episodes can be constructed out of more basic, more flickering memory capacities which do not presuppose self-consciousness.

Descartes the Neurophilosopher

I thank all three reviewers for their favourable remarks on my long rereading of Descartes' fluid-based physiological psychology. Both in his specific associative model of memory and in the remarkable internal complexity and activity which his Cartesian automata exhibit, Descartes' influence and significance is much stranger than our tired textbooks admit, too often simply reinforcing as they do his talismanic place as the demonic source of modern alienation. Meyering in particular is keen to support my naturalistic historical stance in interpreting Descartes as a natural philosopher exploring the limits and the consequences of a mechanistic approach to the various special sciences, rather than as a metaphysician and foundational epistemologist with a merely passing interest in the nature of passive matter. But his acceptance of my 'skewed historical slant' is only partial, for he charges that my interpretation of Descartes as neurophilosopher leaves Descartes' 'confident' and 'exuberant' dualist metaphysics mysterious.

I have two preliminary remarks on Meyering's entirely sensible challenge. First, I don't actually bypass consideration of Descartes' remarks on an 'intellectual memory' which "is certainly independent of the body". Indeed I think my seven-page discussion of the topic offers a fairly complete treatment of Descartes' inconsistent and vague scattered references to this alternative non-physical memory: I reject a chronological story by which Descartes gradually *replaced* a prior corporeal model with this intellectual memory, and I examine the quite different contexts in which he mentions it, in discussions of the resurrection, of infantile amnesia, and of the physiology of wonder (compare Joyce [1997], which I hadn't seen at the time of writing the book). So my interpretation, harping on the incessant motion of bodies and of animal-spirits, is meant at least to address every strand of Descartes' thinking on memory.

But, secondly, there is of course a sense in which it's historically empty to shout "Descartes the connectionist", and in which my reversal of traditional judgements that Cartesian neuromechanics is simply absurd smacks mainly of a corrective mischief. So I'd be happy if attention to Descartes' philosophy of the brain, of imagination and dreaming, passion and memory, was simply integrated more centrally into the growing body of work on Cartesian natural philosophy. I'm not sure what weight to give to Meyering's rhetorical invocation of Descartes' 'rigor and clarity' as a thinker: in my view, it's not to denigrate Descartes to point out that he may have believed certain doctrines with greater confidence at some times and in some contexts than at others. As Catherine Wilson

(forthcoming) argues, one doesn't have to follow La Mettrie in making Descartes a consistent closet materialist to acknowledge that when under less pressure of controversy, Descartes did tend to push his theories of brain and body much further towards an explanation of higher cognitive function.

In particular, I stand by the claim that Descartes saw the difficulties of homuncular explanation, and resisted the notion that perception, memory, and imagination require ideas or brain traces to be inspected in an internal theatre, as if there were 'yet other eyes in our brain', as he mockingly puts it in the *Dioptrics*. So when Wilson writes that, for Descartes, patterns of activation 'were read or experienced' as having content, I'd prefer to say that the reconstruction of the pattern of animal spirit flow simply *is* the (corporeal) remembering. Meyering quite fairly takes to task my pseudo-psychological speculation on the reasons why Descartes still retained a kind of central neurological executive in the form of the pineal gland. This rolling gland mediated a range of complex yet still 'mechanical' responses, which could be delayed over long periods as in (corporeal) remembering, by calling on the resources of the memory traces distributed across the 'folds of the brain'. In Meyering's view, such responses, which we share with beast-machines, shouldn't be taken as evidence within Descartes' scheme of any genuine kind of 'cognitive control' within the physical realm, for genuine control is reserved only for true actions caused by the soul.

Problems of psychological control are, I agree, the key issue here, rather than the usual questions about the difficulties of causal interaction across the metaphysical mind-body divide. Descartes worried directly over such problems of what we'd call autobiographical memory, in trying to understand how it is 'that past things sometimes return to thought as if by chance and without the memory of them being excited by any object impinging on the senses'. What's striking about his approach to such questions of cognitive discipline is that he didn't think that the soul, for all its metaphysical freedom, could simply erase or even easily manage the ongoing dynamics of corporeal memory traces; and that neither did he, like many of his critics, seek out alternative theories of corporeal memory which might minimise or even rule out in advance the very confusions and combinations of ideas in memory which endanger the dominion of reason. Meyering thinks that my remarks on the psychophysiological basis of self-mastery in Descartes might alleviate these tensions between neuro-philosophy and dualism. I agree with his diagnosis of the residual role of the soul in 'unlearning' the physiological habits with which nature, experience, and education has marked our brains: a central message of the *Passions of the Soul* is that we (uniquely) can, gradually and with some

psychological 'work' *(industrie)*, come to apply intelligence even to the reflexes. To add to the exposition and defence of a Cartesian version of distributed memory, then, this is my only meagre resolution of the new problem of the unity of body and mind in the compound living creature: that self-knowledge, specifically including knowledge of my own body, can for Descartes allow the active mind slowly to mould associative responses, becoming an architect of one's passions and of the corresponding landscape of pores and fibres which the animal-spirits sculpt.

'Bound to Words'

Let me turn, finally and more briefly than I would like, to Michael Mascuch's fascinating proposals. Mascuch picks on the difficult task of diagnosing the recurrent retreat of modern memory theorists to over-static localist or archival models of internal storage. Why has the modern individualists' attachment to strong notions of autonomy and responsibility so often been coupled with a theoretical commitment to the possibility of 'total recall', of complete control over items stacked singly in internal memory rooms? Mascuch is uneasy with my casual invocation of ideological or social-psychological factors in explaining the specific resistance of English Restoration natural philosophers like Joseph Glanvill to the reconstructive Cartesian model of memory. I gesture towards a parallel between English desires to order the collective political past after 1660 and the simultaneous attacks (by Glanvill and More, Hooke and Boyle) on theories which threatened the possibility of disorder in the cognitive realm. Mascuch wants, in contrast, to attribute these men's horror at the 'disorderly floating' of images in memory instead to dramatic changes in print culture and to a new stress among natural philosophers on the cognitive utility of 'plain simple clear and uncompounded' representations.

I have no doubt that there is a measure of truth in this diagnosis, and enjoy Mascuch's lovely notion that the graphic, discrete representation of a louse in Hooke's *Micrographia* might have had a more powerful psychological impact than a physical louse roaming Glanvill's scalp. I have a couple of historical observations before commenting on the wider significance of the ambitious style of Mascuch's analysis. Firstly, I suspect that Glanvill's charged attacks on Descartes' (and Hobbes') theories of memory may have had quite different psychological sources than were driving Hooke's localist theory of memory. Where Hooke, working at the heart of the Royal Society, was caught up in problems about the consequences of superposition in the theory of light as well as in memory, and by his own intricate schemes for constructing external textual and

graphic supplements to 'natural' memory, Glanvill's reflections on human cognitive limitations did appear in a straightforwardly religious context of meditation on the aftermath of the Fall. Secondly, I'm not yet persuaded by Mascuch's denial that memory was often modelled on writing by early modern theorists: in a long and mixed tradition stretching from medieval works on the arts of memory through the seventeenth century, the notion of memory as inner writing was a common way of binding the mind to words, rendering the innards static, and internalising even the imperfectly stable artefacts of text in order to maintain the illusion of authorial control over the dangerous contents of one's memory (Sutton 2000).

This example illustrates the difficulty as well as the interest of this kind of 'historical cognitive science', in which specific case studies are sought with which to examine the range of possible interactions between external information-storage systems, and the various formats of mental representation. There is an inevitable looseness of explanatory fit between the history of technological and social representations and the history of memory representations, for acute awareness of the need for precision and rigidity in external representation could just as well spring from acceptance of the weakness and confusion of 'natural' internal memory (as indeed it did for some medieval natural philosophers), as from the opposite sense which Glanvill and his ilk had that an exactly equivalent, context-independent freezing *must* be the true nature of (cognitive) memory. So I don't see that the developments in print culture which Mascuch economically outlines *dictated* Glanvill's imposition of natural order on memory any more than did the political context to which I referred.

But the uncertainty of this form of explanation is no reason not to attempt it. Indeed, it's a burden imposed by reconstructive models of mind that a mature connectionist-inspired cognitive science must *include* attention to changing cultural representational formats. If complex, stable structure isn't built in to a permanent archival system of representations of the world, but can only be constructed out of flickering, narcissistic, context-dependent snapshots, then the optimistic quest for a science of the embodied and extended mind will have to span cognitive anthropology as well as neurobiology, history as well as computational modelling, as we seek to understand how cognitive engines in the wild lean on and parasitise the representational resources of culture and technology.

Department of Philosophy,
Macquarie University,
Sydney, Australia.

References

Campbell, J. (1997) "The Structure of Time in Autobiographical Memory", *European Journal of Philosophy, 5.*

Carruthers, M. (1990) *The Book of Memory: A Study of Memory in Medieval Culture.* Cambridge: Cambridge University Press.

Changeux, J.-P. (1985) *Neuronal Man: The Biology of Mind.* New York: Pantheon Books.

Clarke, E. (1968) "The Doctrine of the Hollow Nerve in the Seventeenth and Eighteenth Centuries", in Stevenson, L. G. & Multhauf, R. P. (eds.), *Medicine, Science, and Culture.* Baltimore: Johns Hopkins University Press.

Churchland, P. M. and P. S. Churchland, (1996) "The Future of Psychology, Folk and Scientific", in R. N. McCauley (ed.) *The Churchlands and their Critics.* Oxford: Blackwell.

Engel, S. (1999) *Context is Everything: the Nature of Memory.* London: W.H. Freeman.

Fine, A. (1975) "How to Compare Theories: Reference and Change", *Noûs, 9.*

Harrington, A. (1987) *Medicine, Mind, and the Double Brain.* Princeton: Princeton University Press.

Hooke, Robert (1665) *Micrographia: or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses.* London.

Joyce, R. (1997) "Cartesian Memory", *Journal of the History of Philosophy, 35.*

Kitcher, P. (1978) "Theories, Theorists, and Theoretical Change", *Philosophical Review, 87.*

Kitcher, P. (1993) *The Advancement of Science.* Oxford: Oxford University Press.

Krell, D. F. (1990) *Of Memory, Reminiscence, and Writing: On the Verge.* Bloomington, Indiana: Indiana University Press.

Locke, John (1690) *An Essay Concerning Human Understanding.* London.

Meyering, T. C. (1989) *Historical Roots of Cognitive Science: The Rise of a Cognitive Theory of Perception from Antiquity to the Nineteenth Century.* Dordrecht: Kluwer Academic Publishers.

Pepys, Samuel, (1970–83) *The Diary of Samuel Pepys.* Edited by Robert Latham and William Matthews. 11 volumes. Berkeley: The University of California Press.

Pitt, Moses (1681) *The English Atlas.* London.

Quine, W. V. O. (1953) "Two Dogmas of Empiricism", in *From a Logical Point of View.* New York: Harper & Row.

Quine, W. V. O. (1971,) "Epistemology Naturalized", presented at the XIVth International Congress of Philosophy, Vienna.

Schacter, Daniel L. (1999) "The Seven Sins of Memory", *American Psychologist, 54.*

Shapin, S. and S. Schaffer (1985) *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life.* Princeton: Princeton University Press.

Sutton, J. (2000) "Body, Mind, and Order: Local Memory and the Control of Mental Representations in Medieval and Renaissance Sciences of Self", in G. Freeland and A. Corones (eds.), *1543 And All That: Word and Image in the Proto-Scientific Revolution.* Dordrecht: Kluwer.

Wilson, C. (forthcoming, 2000) "Descartes and the Corporeal Mind: Some Implications of the Regius Affair", in S. Gaukroger, J. Schuster, and J. Sutton (eds.), *Descartes' Natural Philosophy*. London: Routledge