The Mind: Embodied, Embedded, but not Extended

Gerard O'Brien

Department of Philosophy
University of Adelaide
South Australia 5005

gerard.obrien@adelaide.edu.au
http://arts.adelaide.edu.au/Philosophy/gobrien.htm

Draft @ August 1997

Appeared in Review Symposium on Andy Clark's *Being There: Putting Brain, Body and World Together Again*, Metascience **7:**78-83 (1998)

Perhaps it's a mark of the sheer vitality of the relatively young field of cognitive science that it's grappling with its third major paradigm in the space of just thirty years. While the roots of the discipline can be traced back to 1960s, its real beginnings occurred in the early 1970s with the application of ideas derived from conventional digital computers to human cognition, spawning the now appropriately named *classical* computational theory of mind: the doctrine that cognition is a species of symbol manipulation. Then, in the mid-1980s, the field witnessed its first major shake-up with the advent of neurally-inspired, parallel distributed processing (PDP) computational models, which substituted operations over activation patterns for symbol manipulations, and many theorists in the field started talking passionately about *connectionism*. Now, scarcely ten years later, the field is once again in tumult, this time with the arrival of *dynamical systems theory*, which, because it eschews the concept of representation, threatens to create an even greater rift in the field than that which occurred between connectionism and classicism.

It is in this revolutionary milieu that Andy Clark's latest book *Being There* is situated. Clark rose to prominence through his advocacy of connectionism, with his two previous books (*Microcognition* Cambridge, MA: MIT Press, 1989, and *Associative Engines* Cambridge, MA: MIT Press, 1993) containing some of the most penetrating philosophical work to be found on this alternative approach to the mind. But Clark, who might have expected to spend a few more years developing connectionism in a relatively stable intellectual environment, now finds himself defending it against the even newer dynamical systems vision of cognition.

Clark's response to this predicament is to preach ecumenism. Just as *Microcognition* argued that we shouldn't throw out all the classical insights as we stampede towards connectionism, *Being There* puts the case for combining the embodied, embedded aspects of cognition highlighted by dynamical models, with the commitment to representation, and hence computation, that we find in connectionism (and classicism, for that matter). This is a sensible position, in my view. And there is much to admire in Clark's latest book. He is a gifted expositor, and *Being There* is brimming with detailed and entertaining discussions of the new light that dynamical systems theory is throwing on the role played by both the body and the environment in shaping cognitive processes. At the same time he doesn't shy away from providing incisive critiques of the excesses of this programme, especially when these bubble over into what Clark terms the "Thesis of Radical Embodied Cognition", the claim that "embodied cognition is best studied by means of noncomputational and nonrepresentational ideas and explanatory schemes" (p.148). His point here is that such radicalism is unjustified and

counter-productive, inviting competition between dynamical and computational conceptions of cognition where progress is more likely to be achieved through cooperation (see especially chp.8).

While there is much in *Being There* that I like, the very nature of review symposia forces the commentator to look for points of discord rather than concurrence, simply because disagreements are bound to more interesting and response-provoking. In what follows, therefore, I will focus on the one major ecumenical theme propounded in *Being There* that I find difficult to accept. This is Clark's advocacy, especially in the third and final part of the book, of the *extended* nature of the embedded, embodied mind.

Talk of the mind leaking out of the brain and into the world is in the air these days. In philosophy it's primarily driven by externalist theories of mental content, which hold that the meaning of some mental states is determined by the causal relations that internal, brain states bear to extrinsic, environmental factors. But Clark is quite explicit that his motivation is quite different (see especially fn.23, p.246). For him the seepage of the mind into the environment is licensed by the subtle couplings between the brain and aspects of the environment, so emphasised by dynamical systems theory, that make it reasonable to suppose that certain extra-bodily resources play a *constitutive* role in some cognitive operations. This kind of extension is most plausible, he thinks, "in cases involving the external props of written text and spoken words, for interactions with these external media are ubiquitous..., reliable, and developmentally basic" (p.214). And his conclusion is that in such cases, "what we commonly identify as mental capacities may...turn out to be properties of the wider, environmentally extended systems of which human brains are just one (important) part" (p.214).

Clark is well aware that, without qualification, this thesis is in danger of foundering on the reef of common sense – the distinction between my mind and yours should not be allowed to collapse "just because we are found chattering of the bus" (p.217). So there must be principled ways of isolating those external props that become part of the mind from the absolutely vast number that don't. Some of the constraints Clark suggests here are that the requisite information must be "easy to access and use", "automatically endorsed", and "originally gathered...by the current user" (p.217). His favourite example is that of a notebook, which is our constant companion, and in which we make all manner of scribblings. The crucial point in such a case, he argues, is that "the entries in the notebook play the same explanatory role, with respect to the agent's behavior, as would a piece of information encoded in long-term memory" (p.218). It is principally this "functional isomorphism" that licenses his contention that our "beliefs, knowledge, and perhaps other mental states now depend on physical vehicles that can (at times) spread out to include select aspects of the local environment" (p.218).

In making these claims about the mind's extension beyond the skin and skull, Clark is opting for one of the two traditional ways of distinguishing between the mental and the merely physical. One way is to suppose that *consciousness* is the mark of the mental, and hence determines the extent of the mind. But Clark thinks that conscious experience is fully explained by the current state of the brain, so there is no basis here for any mind expansion (see pp.216-7). The other way, the way Clark relies on, is to focus on the property of *intentionality*, whereby mental states possess the property of aboutness or, in the language of cognitive science, representational content. The mind's boundaries, according to this second approach, are drawn around the representational vehicles it manipulates in the course of cognition. And so intimate is the causal commerce between human brains and certain written and spoken words, according to Clark, that these external artefacts themselves constitute part of the mind's representational substrate.

But I'm not convinced. It's not that I object to the general criterion by which Clark seeks to include these representational vehicles in the mind (namely, that they are functionally

isomorphic with the those that standardly encode information in long-term memory). The problem, as I see it, is that, at least in the context of a broadly connectionist understanding of cognition, even his best examples fail to satisfy this condition. No matter how vigorous the causal commerce between parts of my mind and information I record in a personal notebook, these external symbols do not have the same causal properties as the representational vehicles responsible for my memories. To see this, it's necessary to step back somewhat and rehearse some of the now fairly familiar details of the mind's information coding and processing capacities, as these are understood from a connectionist perspective.

It's commonplace for theorists to distinguish between *explicit* and *nonexplicit* forms of information coding in a computational device. Representation is typically said to be explicit if each distinct item of information in the device is encoded by a physically discrete object. Information that is either stored dispositionally or embodied in a device's primitive computational operations, on the other hand, is said to be nonexplicitly represented. It is reasonable to conjecture that the brain employs these different styles of representation. Connectionists make much of this distinction by pointing to the two different ways in which information is coded in PDP networks, and hence by extension, in the brain's neural networks.

The representational capacities of PDP systems rely on the plasticity of the connection weights between their constituent processing units. By altering these connection weights, one alters the activation patterns the network produces in response to its inputs As a consequence, an individual network can be taught to generate a range of stable target patterns in response to a range of inputs. These stable patterns of activation are semantically evaluable, and hence constitute a form of information coding. What is more, because these patterns are physically discrete, structurally complex objects, which each possess a single semantic value, it is reasonable to regard the information they encode as *explicitly* represented.

While activation patterns are a transient feature of PDP systems, a "trained" network has the capacity to generate a whole range of activation patterns, in response to cueing inputs. So a network, in virtue of its connection weights and pattern of connectivity, can be said to *store* appropriate responses to input. This form of information coding constitutes long-term memory in PDP systems. Such long-term storage of information is *superpositional* in nature, since *each* connection weight contributes to the storage of *every* stable activation pattern (every explicit representation) that the network is capable of generating. Consequently, the information that is stored in a PDP network is not encoded in a physically discrete manner. The one appropriately configured network encodes a *set* of contents corresponding to the range of explicit tokens it is disposed to generate. For all these reasons, a PDP network is best understood as storing information in a *nonexplicit* fashion.

These facts about information coding in PDP systems have major consequences for the manner in which connectionists conceptualise cognitive processes. Most importantly, information that is nonexplicitly represented in PDP networks need not be rendered explicit in order to be causally efficacious. This is because it is a network's connection weights and connectivity structure that is responsible for the manner in which it responds to input (by relaxing into a stable pattern of activation), and hence the manner in which it processes information. There is a strong sense, therefore, in which it is the nonexplicit information in a network (i.e., the network's "memory") that governs its computational operations: *all* the information that is encoded in this fashion is causally active *whenever* that network responds to an input. The causally holistic nature of information processing in PDP systems is the reason that many theorists think that connectionism provides us with a hint as to how Nature might have solved the infamous frame problem. From a connectionist perspective it's possible to envisage how, whenever we act in the world, a very large amount of information could be automatically and unconsciously guiding our behaviour.

But these same facts about connectionism would appear to be destructive of Clark's attempts to extend the mind beyond the skull. It is quite clear that the information encoded in the form of symbols in a personal notebook doesn't have these causal properties, and hence isn't functionally isomorphic with the information contained in our long-term memory. There are two important points of difference here. The first is that the external symbols are *causally passive*: the information they encode doesn't do any work unless we bring them under the gaze of our perceptual equipment. At this point the recorded information does become causally active, but only because it is now *re-coded* elsewhere – namely, inside our skulls. The second difference is that such externally recorded information, when it does become causally engaged with parts of the mind, does so only in a *causally discrete* fashion: each separate piece of information, coded by a distinct symbol structure, must be individually accessed and processed. These differences would thus seem to mark an important natural boundary; one that makes it hard to justify, even on Clark's own terms, any extension of the mind's representational substrate to include our written and spoken words.

Incidentally, this talk of the causal passivity and discreteness of external symbols should call to mind one of the of-cited differences between connectionism and classicism. One of the reasons that classicism presents a very different picture of cognition from connectionism is because it holds that information in long-term memory, unlike the information stored in a PDP network, is just like the information recorded on a piece of paper, in that it must be discretely accessed by some processing mechanism before it can causally influence ongoing cognitive operations. (This is not to say that classicists are committed to the view that all long-term memories are stored explicitly. In fact, given the sheer bulk of information that is stored in the brain, classicists are committed to the existence of highly efficient, generative systems of information storage and retrieval, whereby most of our knowledge can be readily derived, when required. But such information, while stored in an nonexplicit form, must first be rendered explicit before it can be causally effective.) So Clark's case for extending the mind across those symbols inscribed in various external media would be much stronger in the classical context. But this just serves to highlight why it is a mistake to enlarge the mind's boundaries in this way. As many theorists have argued, it is precisely because classicism is committed to this account of memory and information processing (i.e., because it is committed to the code/process divide – see, e.g., Clark's *Associative Engines*) that the infamous "frame problem" – the problem of equipping a cognitive system with the wherewithal to choose appropriate courses of action, in response to internal goals and changing environmental conditions, in real time – is so acute for this approach to cognition.