symbolic 'props and pivots' of a form which is, in a sense, profoundly unnatural. Like the medieval monks who laboriously forced strange architectural memory palaces into their minds so as to keep stored items distinct, to guarantee immunity from the melding characteristic of 'natural' memory, we all impose (an approximation of) rigidity and inflexibility on our own mental representations. As the dolphins teach us, of course, supplements to our bare biology are responsible for many wonderful extensions to our capacities: but Clark's stress on the generality of (at least some of) our learning mechanisms reminds us that the specific cognitive trajectories along which our particular cultural and institutional learning aids allow us to go are, in a way, deeply contingent. Clark's version of dynamical cognitive science foregrounds the action-oriented and path-dependent nature of 'mind on the hoof' (p. 35), and it opens up vast theoretical terrain in which it may be possible to attend to brains and contexts at once.

School of Philosophy, University of Sydney,
Sydney, New South Wales,
Australia

Author's Response

By Andy Clarke

THERE can be few things more satisfying than reading friendly, constructive engagements with one's own work.[1] I thank the four reviewers for their patient and penetrating comments, and for the truly marvellous overviews of the project. The pieces by Hooker and Sutton distil the essence of the project with great and enviable clarity, while all four reviewers push, probe and extend the work in challenging yet helpful ways.

The general idea of *Being There* was to weave a variety of sometimes unlikely looking components into a coherent (but somewhat non-standard) view of natural intelligence: a view in which basic organism/environment coupling is fundamental and in which advanced cognition emerges as deeply continuous with these roots. A major element of the story, as noted by several reviewers, was a highly generalised notion of 'scaffolding'—of bodily and environmental structures (including linguistic and cultural artefacts) that re-shape the space of individual reason and thus enable us to press maximal benefit from fragmentary, pattern-completion styles of internal computational organisation.

Such a view, although not mainstream, is certainly not novel. Hooker's own work on control theory, the very substantial literatures of 'new robotics', artificial life and dynamical systems theory, and the more philosophical frameworks of Varela, Lakoff, Johnson and others, are all clear examples of closely related views mentioned in the text. Work in connectionism, cognitive anthropology, education and economics is also invoked and a major goal of the book was to try to coax these various elements together to isolate some unifying themes, and to highlight some problematic issues.

The coaxing together seems to have been largely successful, and the reviewers' appreciative comments warmed my heart on a cold morning. One reviewer (Quinn) goes on to suggest an interesting extension to the set of core elements—a proposal I will return to later. For the most part, however, the critical comments focused on three of the more troublesome issues raised by the text. First, the unexplicated notion of agent autonomy; second, the problematic suggestion that mind might somehow leak out into the surrounding world; and third, the vexed role of internal representation in the explanation of intelligent behaviour. I shall take these in turn, then end by discussing Quinn's proposed extension and some possible future developments.

Autonomy

Cliff Hooker's stylish and engaging commentary highlights an important question—one that is, I confess, not even addressed in the book. I make extensive use of the popular term 'autonomous agent' but say nothing about the nature of the autonomy itself. Worse still, the examples I give of real-world artificial 'autonomous agents' are, Hooker suggests, not really autonomous agents at all, although they do "share some of the same general functional features as autonomous systems".

Hooker's view, as I understand it, is that genuine autonomy involves a special kind of intelligent control of action, what he calls adaptable, anticipative control. Autonomous, Adaptable, Anticipative Systems (AAA Systems) are ones that modify their own responses and routines so as to create and sustain a life (or functionality) preserving coupling with their environments. A robot such as Herbert (the soda can collecting robot described in the early pages of *Being There*) is not an AAA System, as its activity is not adaptably geared to maintaining its own functionality. AAA Systems, Hooker suggests, display a type of organisation that goes beyond "mere dynamical pattern formation". If we identify cognitive systems as AAA Systems, then we can see, rather concretely, why cognition involves a special kind of agent–environment coupling.

This strikes me as a good way to go. The strong sense of autonomy that Hooker defines does allow us to mark some important discontinuities in the design space that is being explored by contemporary work in robotics and artificial life. My own guess, however, is that the notion of anticipative, adaptable response is itself still too broad and disunified to mark any rigid boundary between cognitive and non-cognitive routes to adaptive success. Indeed, part of the thrust of *Being There* is to suggest that the cognitive/non-cognitive distinction is itself too coarse a tool to bear real scientific weight. Certain kinds of simple insects and maybe even some plants may well fit the basic image of an AAA System, exhibiting both some degree of learning and of self-modification geared to survival. What we will probably find then (and I have no reason to think that Hooker disagrees with this) is that a lot depends on the different ways in which anticipative, adaptable response is supported. (In a later section, I will comment on one such way: the use of inner circuits to emulate agent/environment dynamics).

On the topic of autonomy, I would also flag Tim Smither's interesting work (e.g., Smithers, ms) which seems to dovetail nicely with Cliff Hooker's. Smithers argues that true autonomy requires a process of "self law-making", not just self-regulation. An example of this would be systems which actively create the kinds of environment (both internal and external) they need in order to function efficiently. Such a notion of autonomy also fits well with the observation, central to *Being There*, that intelligent behaviour often depends on the creation and exploitation of 'external scaffolding'—environmental structures that simplify and reconfigure the tasks confronting biological brains.

In sum, I agree that *Being There* works with a broad and unanalysed notion of "autonomous agent". In my defence, I note that so do most real-world robotics laboratories and that the broad notion (of embodied, usually mobile devices capable of simple real-world real-time activity) does pick out an interesting class of systems. But I agree that a stronger notion of autonomy may help identify important discontinuities in design space (see Sloman 1994). And much of my current work is indeed concerned to fine-tune the story in just these kinds of way (see especially Clark, in press; Clark and Grush, submitted).

Seepage

Gerard O'Brien approaches me from a different angle, with a deft blow to an acknowledged weak spot: the consistency organ. O'Brien worries about the idea (pursued gently in the book and more vigorously in Clark and Chalmers 1995) that mind may sometimes seep outside the traditional envelope of skin and skull, inhering instead in extended systems comprising the biological brain and selected aspects of the body and local environment. The reason why this doesn't happen, he argues, lies in the different ways in which external and internal components store and organise information: differences that ought to have been especially clear to the author of two books (Clark 1989, 1993) contrasting connectionist and classical modes of information storage and retrieval. (Hence the threat to the consistency organ.)

More precisely, O'Brien argues that external information stores (such as the well-maintained and constantly available notebook featured in Chapter 10 of the book and in Clark and Chalmers 1995) are not plausibly seen as functionally isomorphic to biological long-term memory, at least as depicted by connectionist theory. Such a notebook might indeed be somewhat similar to a classical vision of an inner data-base. But the connectionist vision, with its stress on superpositional information storage (and on associated properties such as free generalisation, content addressability and graceful degradation) paints a quite different picture. If the connectionist story is (as it seems to be) closer to the natural facts than the classical one, then there is indeed a world of difference between the passive discrete symbol structures found in the typical external store and the active inexplicit representations found in the head.

O'Brien depicts my suggestion that mind might seep out into the world as based entirely on a principle of functional isomorphism: if some element outside the head is contributing to behavioural success in a way that is functionally isomorphic to the contribution of some inner, standardly cognitive resources, then it should be seen as part of the cognitive system too. But I think he reads too much into the (perhaps ill-advised) locution of 'functional isomorphism'. For the isomorphism is said to hold only in respect of the explanatory role of the external elements in a commonsense account of the agent's behaviour. The basic idea (developed more fully in Clark and Chalmers 1995) is that the notebook entries explain the same kinds of very broad patterns of purposive behaviour as does knowledge stored in biological memory. To that, O'Brien will reply (I suppose) that the kinds of pattern provided for are really subtly different, perhaps in respect of properties such as generalisation and the like. To which we will reply that these differences leave intact a more fundamental similarity concerning the appeal to stored information in the explanation of purposive action.

Such an exchange, however, only gets us so far. A better response to O'Brien's critique is, I think, to see it as identifying a potential tension between two components of the extended mind story itself. One component (the one he focuses on) stresses the way that extra-neural elements can play a role similar to internal ones (as in talk of external memory, etc.). But a second component, which was repeatedly highlighted in the text, turned on the way external elements may play a role different from, but complementary to, the inner ones. It is this vision that is invoked in the discussion of Hutchins' work on the role of maps, compasses and so on in an extended (multi-agent and artefact) ship navigation system: a discussion I explicitly cite (p. 214) in introducing the topic of the extended mind. This same complementarity is foregrounded by the claim that the user–artefact relationship may be as close and intimate as that of the spider and the web (p. 218) and by the analogy (ch. 11) with the tuna's active creation of water-bound eddies and vortices so as to improve its aquatic performance.

Given this second line of argument (the one stressing complementarity), it is best to see functional isomorphism as at most part of a sufficient condition for cognitive extension, rather than as a necessary feature. The more interesting and plausible argument, I feel, is the one which describes the seepage of mind into the world by stressing that "the brain's brief is to provide complementary facilities that will support the repeated exploitation of operations upon the world [and] to provide computational processes (such as powerful pattern completion) that the world, even as manipulated by us, does not usually afford" (*Being There*, p. 68).

It should be clear enough, from this last quote, that I have certainly not forgotten the lessons that connectionism taught us. The argument for the extended mind thus turns primarily on the way disparate inner and outer components may co-operate so as to yield integrated larger systems capable of supporting various (often quite advanced) forms of adaptive success. The external factors and operations, in this model, are most unlikely to be computationally identical to the ones supported directly in the wetware—indeed, the power of the larger system depends very much on the new kinds of storage, retrieval and transformation made possible by the use of extra-neural resources (see also the tale of John's Brain told in the appendix). These new operations, however, may often be seen as performing kinds of tasks which, were they but done in the head, we would have no hesitation in labelling cognitive. This is because they contribute to behavioural success by for example storing and manipulating information, and by reconfiguring problem spaces. This kind of higherlevel functional isomorphism is, I think, quite compatible with the idea (stressed by both O'Brien and myself) that there exist deep and important differences between e.g., active biological and passive symbolic modes of storage and retrieval.

Representation (and computation)

Both Sutton and Hooker would like to see a more fully worked-out story about how to factor internal representation and computation into the larger, ecumenical package of *Being There.* So would I. As it stands, the chapter that tackles these topics (Chapter 8: "Being, Computing Representing") is both the largest and the most frustratingly 'unfinished' one in the book. In it, I argue for what I call 'minimal representationalism': the view that we need to combine dynamical and ecological analyses with the search for in-the-head states and processes that both encode contents (albeit, often fragmentary, action-specific kinds of content) and that exploit computational routines so as to systematically transform one content into another. Such states and processes, I argue, are most strongly implicated in episodes in which we reason about absent, counterfactual or imaginary states of affairs.

Sutton queries the point about thoughts concerning the absent. Instead of persisting inner surrogates for what is not present-to-hand, Sutton proposes that we create such surrogates on the spot, out of the whole cloth of a complex web of inner and outer dynamics. But I have no problem with such an account. All it means (if true) is that the inner surrogate comes into being as and when it is needed. This is fine by me: what matters is (still) that on-going behaviour, in such cases, is explained by appeal to identifiable inner content-bearers. The stability and long-term persistence of such items is not an issue on which I have to take a stand.

That said, I should concede the more general substance of Sutton's worry. For it is true that it is not inconceivable that complex, evolving inner states, of some kind which does not succumb to any fine-grained content-ascribing decomposition, might somehow support behaviour which is coordinated with respect to distal, absent or non-existent states of affairs. We cannot rule this out *a priori*, and some researchers in Artificial Life and real-world robotics are already trying to solve such coordination problems without making any prior commitments to the use of internal representation (e.g., Beer 1996).

My own view, however, is that the most practical and efficient mechanisms for coordinating complex behaviour with what is absent, imaginary and counterfactual will involve the use of systems of inner states and processes whose functional role is to stand-in for the 'missing' states of affairs—in short, internal models and internal representations. In recent (post-*Being There*) work, I have pursued this idea using some of the apparatus mentioned by Hooker who asks "could off-line emulation be the intended source of Clark's representation?". Very briefly, the idea (pursued at length in Clark and Grush, submitted; and also in Clark, in press) is that internal representation, strongly conceived, gets its foot in the door of biological cognition when on-line, real-time behaviour requires a system to adjust certain parameters on the basis of information that is not available fast enough to allow direct control by environmental feedback. It is speculated, for example (see Ito 1984, Kawato *et al.* 1987, Dean *et al.* 1994) that the control of reaching requires proprioceptive feedback to be deployed before real signals from the sensory peripheries could be exploited. A solution is to train on-board circuitry to mimic the dynamics of the larger system and to generate a prediction of the real signal that can then be used to fine-tune the reaching. The emulator circuit thus acts as a stand-in for the real-world system itself. Although I mention this work in the book (pp. 22–3), it is not there developed into a general story about (strong) internal representation. The development (again, see Clark and Grush, submitted) involves noting that such an emulator, though originally invoked to fine-tune actual reaching, may be run off-line so as support motor imagery without real-world action (see Grush 1995). In such cases we can actively isolate the precise aspects of the processing that correspond to different target events and states of affairs (in the reaching case, to different arm motion parameters). Our suggestion is that a creature uses full-blooded internal representations if and only if it is possible to identify within them specific states or processes whose functional role is to act as de-coupleable surrogates for specifiable (usually extra-neural) states of affairs.[2] Motor emulation circuitry, we think, provides a clear, minimal and evolutionarily plausible case in which these conditions are met. And it is shows how internal representations might first originate in systems whose 'goal' is merely to maintain close and fluent behavioural contact with the world around them.

The Future

Naomi Quinn, in her richly suggestive and multi-layered commentary, offers a fascinating counterpoint to my tendency to depict cultural scaffolding as external and as heavily linguistic. Quinn's emphasis, by contrast, is on the "unspoken, internal cultural representations that mediate performance of . . . cognitive tasks". These involve, as I understand it, shared culture-specific ideas and metaphors that, although often unconscious and unarticulated, serve to structure our understanding, judgement and responses. Quinn depicts, in persuasive detail, the content of (to take one example) a shared cultural representation of marriage as a lasting, yet fundamentally contractual and mutually beneficial, relationship. Such shared conceptions make it possible to construct arguments and discourses whose flow depends crucially on unstated, invisible premisses and assumptions. The presence of such a shared backdrop reduces cognitive load and scaffolds problem-solving: yet the scaffolding consists neither in external structures nor in linguistic productions, inscriptions or rehearsals.

I think Quinn is right to depict this as a kind of cognitive scaffolding and as a way in which culture seeps into the mind. Such internal scaffolding helps to enforce a kind of mental hygiene by both restricting and propelling our reasoning and inference. (Sutton's lovely description of the role of linguistic rehearsal has a natural extension to this kind of unarticulated, schematic case: the culturally inherited schemes act as a kind of pivot for linguistic and interpersonal reason.)

My only fear, in all this, is that the notion of scaffolding could one day grow too broad. It would not do, for example, if every aspect of cognition could be seen as performing a scaffolding function. We need to maintain a sense that the scaffolding involves elements that are in some hard-to-pin-down sense external to the most basic processes of biological reason. I think, however, that the case of internal cultural representations probably qualifies, insofar as we are there dealing with inner states whose shape, content and role are fixed by some quite specific social and collective practices which seem to reconfigure on-board reason in ways not predictable from a more individualistic stance. But however we describe them, Quinn is surely right to flag an important dimension of analysis ignored in my original treatment.

There are other directions, also, in which I hope to extend the original project. One is to look more closely at the question of biological implementation; to ask whether neural computation might be pressing important functionality out of 'mere implementation details' such as the low-level physics of the hardware (see e.g., Thompson 1996). Another is to look at the 'double life' of beliefs and ideas, being on the one hand mental entities ascribed to individual agents and, on the other hand, entering into larger, collective dynamics that have properties all their own (think of the way ideas and beliefs interact and snowball in financial markets—see Arthur 1997). Accommodating this 'double-aspect' of beliefs and ideas is, I suspect, going to prove crucial to the understanding of many forms of cultural scaffolding. In addition (and as we saw), the respective explanatory roles of dynamics, computation and representation are still somewhat up for grabs. Terms of art such as 'emergence' and 'scaffolding' probably require more work. And the whole issue of the mind's (putative) extension into the world is begging for further work and reflection. So there is plenty to do!

I would like to end, however, on a truly positive note. It has been a striking (and tremendously gratifying) feature of the response to *Being There* that it has found favour amongst a truly wide diversity of disciplines and readers. In particular, I am greatly excited by the response from the social sciences, cultural anthropology, education, business and economics, as well as philosophy and the traditional cognitive sciences. There is, in the current climate, a real opportunity (or so it seems to me) to now draw together a rich, diverse and highly multi-disciplinary base in pursuit of a truly integrated science of the mind: a science that confronts cognition on its home turf, as the activity of social agents locked in the enabling embrace of culture, artefact and world.

Department of Philosophy,
Washington University,
St Louis, Missouri, USA.

[1] I just thought of seven.

[2] It is a nice question whether there is a coherent weaker sense of internal representation applicable to cases where the 'de-coupleability' criterion is not met. For an attempt to pin down such a weaker sense, see Wheeler and Clark (in progress).

References

Arthur, B. (1997) "Beyond rational expectations" in J. Drobak and J. Nye (eds), *The Frontiers of the New Institutional Economics.* London: Academic Press.

Beer, R. (1996) "Towards the evolution of dynamical neural network for minimally cognitive behavior". *Proceedings of the Society for Adaptive Behavior.*

Clark, A. (1989) *Microcognition.* Cambridge, Mass: MIT Press.

Clark, A. (1993) *Associative Engines.* Cambridge, Mass: MIT Press.

Clark, A. (in press) *The Dynamical Challenge.*

Clark, A., and Chalmers, D. (1995) "The Extended Mind". *PNP Research Report.* Washington University, USA.

Clark, A., and Grush, R. (submitted) "Towards a Cognitive Robotics". *Adaptive Behavior.*

Dean, P., Mayhew, J., and Langdon, P. (1994) "Learning and Maintaining Saccadic Accuracy: A Model of Brainstem-Cerebellar Interactions". *Journal of Cognitive Neuroscience.*

Grush, R. (1995) "Emulation and Cognition". PhD Dissertation, University of California.

Ito, M. (1984). *The Cerebellum and Neural Control.* New York: Raven Press.

Kawato, M., Furukawa, K., and Suzuki, R. (1987) "A hierarchical neural network model for the control and learning of voluntary movement". *Biological Cybernetics.*

Sloman, A. (1994) "Explorations in design space". Paper presented at the 11th European Conference on AI (ECAI), Amsterdam.

Smithers, T. (ms) "Autonomy in robots and other agents".

Thompson, A. (1996) "Unconstrained evolution and hard consequences". *Cognitive Science Research Report* (CSRP), University of Sussex, UK.

Wheeler, M. and Clark, A. (in progress) "Genic representation: Reconciling content and causal complexity".