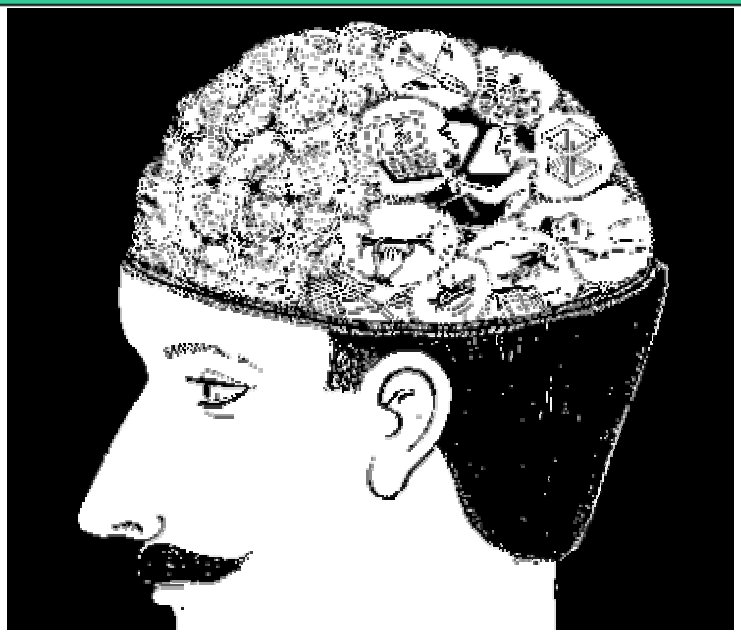
"I am John's Brain" JOURNAL OF CONSCIOUSNESS STUDIES,2:2:1995 p.144-8

Journal of Consciousness Studies
controversies in science and the humanities



Andy Clark:
"I am John's Brain"

The brain and its "agent" debate the provenance of thoughts in the charming language of an old *Readers Digest* article.

Illustration: Jack Buckmaster

I AM JOHN'S BRAIN [1]

Andy Clark [2]
Philosophy/Neuroscience/Psychology Program, Department of Philosophy, Washington University, St. Louis, MO 63130, USA
E-mail: andy@twinearth.wustl.edu

Journal of Consciousness Studies, **2**, No. 2, 1995, pp. 144-8

I am John's[3] brain. In the flesh, I am just a rather undistinguished looking grey/white mass of cells. My surface is heavily convoluted and I am possessed of a fairly differentiated internal structure. John and I are on rather close and intimate terms; indeed, sometimes it is hard to tell us apart. But at times, John takes this intimacy a little too far. When that happens, he gets very confused about my role and functioning. He imagines that I organize and process information in ways which echo his own perspective on the world. In short, he thinks that his thoughts are, in a rather direct sense, my thoughts. There is some truth to this of course. But things are really rather more complicated than John suspects, as I shall try to show.

In the first place, John is congenitally blind to the bulk of my daily activities. At best, he catches occasional glimpses and distorted shadows of my real work. Generally speaking, these fleeting glimpses portray only the products of my vast subterranean activity, rather than the processes which give rise to them. Such products include the play of mental images or the steps in a logical train of thought or flow of ideas.

John's access to these products is, moreover, itself a pretty rough and ready affair. What filters into his conscious awareness is somewhat akin to what gets on to the screen display of a personal computer. In both cases, what is displayed is just a specially tailored summary of the results of certain episodes of internal activity: results for which the user has some particular use. Evolution, after all, would not waste time and money (search & energy) to display to John a faithful record of inner goings-on unless thay could help John to hunt, survive and reproduce. John, as a result, is appraised of the bare minimum of knowledge about my inner activities. All he needs to know is the overall significance of the upshot of a select few of these activities: that part of me is in a state which is associated with the presence of a dangerous predator and that flight is therefore indicated, and other things of that sort. What John gets from me is thus rather like what a driver gets from an electronic dashboard display: information pertaining to those few inner and outer parameters to which the gross activity of the agent can make a useful difference.

John, however, begs to differ. He thinks this is a crazy parallel since in his case there is no further agent to be informed by any `dashboard display'. There is no `driver' apart from me, his brain. But despite this undoubted fact, I insist that there is a dashboard display of sorts. The display consists of those select products of my activities which are able to play a role in those projects and decisions which the world at large ascribes to John-the-person (as opposed to those, like the maintenance of blood flow, ascribed not to John's decisions, but to John-the-biological-organism). The dashboard display thus consists of those products of my activity which are able to figure in what other humans would identify as John's plans, his choices and projects. Thus if one of my many sub-systems is appraised of some item of information, that item may or may not become available to support John's conscious planning and deliberate action. Information which is made available for such purposes can, of course, figure in John's on-going reflections on his own life and goals, while the rest, though often vital for John's continued success, remains invisible to John-the-agent. The fact that John has only limited access to my operations means, of course, that John can sometimes be unaware of the true causes of his own actions. In such cases, John is driven to create complex stories or narratives which try to make sense of his self-observed behaviours. This is a hard task, since the roots of much of that behaviour lie, I am proud to report, in those other activities of mine to which John has no conscious access. As a result, his stories are often wildly imaginative (that is to say, false) attempts to make sense of his own activities on the restricted basis of the `dashboard display' types of information.

And it gets worse. For John's reports, even of the favoured `dashboard display' products of my activity, are themselves filtered through the distorting lens of John's biased and limited vocabulary for reporting these facts to others. Thus John thinks (falsely) that introspection reveals the presence of entities he calls `beliefs', others he calls `desires', still others he calls `hopes' and so on and so on. John is even inclined (in more philosophical moments) to picture these putative inner entities as sharing the basic structure of the very sentences he would use to report such facts to others. He thinks he finds in himself the belief that Rome is pretty, and the

hope that St. Louis is pretty. And just as these sentences share a word `pretty', so John believes the internal states which `carry' the thoughts must share a component too. I do not know why John thinks this, although at times he has such a loose idea of a `component' that what he says cannot help but be true. I assure you, however, that on any non-trivial reading, what he says is false. John should beware of confusing the structure of the language he uses to report his beliefs with the structure of my own encodings. I like to store information in ways which make my unseen labours easier and which come naturally given my evolutionary history — a proud and long one for most of which the recent fad of language-use had not even been invented. My modes of information storage and retrieval, I can safely say, bear no deep resemblance whatever to these new-fangled linguistic vehicles with which John is so misleadingly familiar.

A further complex of misapprehensions centres on the question of the provenance of thoughts. John thinks of me as the point source of the intellectual products he identifies as his thoughts. But, to put it crudely, I do not have John's thoughts. John has John's thoughts and I am just one item in the array of physical events and processes which enable that thinking to occur. John is an agent whose nature is fixed by a complex interplay between a mass of internal goings-on (including my activity) and a particular kind of physical embodiment and a certain embedding in the world. The combination of embodiment and embedding provides for persistent informational and physical couplings bet-ween John and his world; couplings which leave much of John's `knowledge' out in the world and available for retrieval, transformation and use as and when required.

Take a simple example. A few days ago, John sat at his desk and worked rather hard for a sustained period of time. Eventually he got up and left his office, satisfied with his day's work. `My brain', he reflected (for he prides himself on his physicalism), `has done very well. It has come up with some neat ideas.' John's image of the events of the day depicted me as the point source of those ideas; ideas which he thinks he captured on paper as a mere convenience and a hedge against forgetting. I am, of course, grateful that John gives me so much credit. He attributes the finished intellectual products directly to me. But in this case, at least, the credit should be extended a little further. My role in the origination of these intellectual products is certainly a vital one: destroy me and the intellectual productivity will surely cease! But my role is more delicately constituted then John's simple image suggests. Those ideas of which he is so proud did not spring fully formed out of my activity. If truth be told, I acted rather as a mediating factor in some rather complex feedback loops encompassing John and selected chunks of his local environment. Bluntly, I spent the day in a variety of close and complex interactions with a number of external props. Without these, the finished intellectual products would never have taken shape. My role, as best I can recall, was to support John's re-reading of a bunch of old materials and notes, and to react to those materials by producing a few fragmentary ideas and criticisms. These small responses were stored as further marks on paper and in margins. Later on, I played a role in the re-organization of these marks on clean sheets of paper, adding new on-line reactions to the fragmentary ideas. The cycle of reading, responding and external re-organization was repeated again and again. At the end of the day, the `good ideas' (with which John was so quick to credit me) emerged as the fruits of these repeated little interactions between me and the various external media. Credit thus belongs not so much to me as to the spatially and temporally extended process in which I played a role.

On reflection, John would probably agree to this description of my role on that day. But I would caution him that even this can be misleading. For so far I have allowed myself to speak as if I were a unified inner resource contributing to these interactive episodes. This is an illusion which the present literary device encourages and one which John seems to share. But once again, if truth be told, I am not one inner voice but many. I am so many inner voices, in fact, that the metaphor of the inner voice must itself mislead. For it surely suggests inner sub-agencies of some sophistication and perhaps possessed of a rudimentary kind of self-consciousness. In reality, I consist only of multiple mindless streams of highly parallel and often relatively independent computational processes. I am not a mass of little agents so much as a mass of non-agents, tuned and responsive to proprietary inputs and cleverly orchestrated by evolution so as to yield successful purposive behaviour in most daily settings. My single voice, then, is no more than a literary conceit.

At root, John's mistakes are all variations on a single theme. He thinks that I see the world as he does, that I parcel things up as he would, that I think the way he would report his thoughts. None of this is the case. I am not the inner echo of John's conceptualizations. Rather, I am their somewhat alien source. To see just how alien I can be, John need only reflect on some of the rather extraordinary and unexpected ways that damage to brains like me can affect the cognitive profiles of beings like John. Damage to me could, for example, result in the selective impairment of John's capacity to recall the names of small manipulable objects, yet leave unscathed his capacity to name larger-scale ones. The reason for this has to do with my storing and retrieving heavily visually-oriented information in ways distinct from those I deploy for heavily functionally-oriented information; the former mode helps pick out the large-scale items and the latter the small-scale ones. The point, at any rate, is that this facet of my internal organization is altogether alien to John — it respects needs, principles and opportunities of which John is blissfully unaware. Unfortunately, instead of trying to comprehend my modes of information storage in their own terms, John prefers simply to imagine that I organize my knowledge the way he, heavily influenced by the particular words in his language, organizes his. Thus he supposes that I store information in clusters which respect what he calls `concepts' — generally, names which figure in his linguistic classifications of worldly events, states and processes. Here, as usual, John is far too quick to identify my organization with his own perspective. Certainly I store and access bodies of information; bodies which together, if I am functioning normally, support a wide range of successful uses of words and a variety of interactions with the physical and social worlds. But the `concepts' which so occupy John's imagination correspond only to public names for grab-bags of knowledge and abilities whose neural underpinnings are in fact many and various. John's `concepts' do not correspond to anything especially unified as far as I am concerned. And why should they? The situation is rather like that of a person who can build a boat. To speak of the ability to build a boat is to use a simple phrase to ascribe a whole panoply of skills whose cognitive and physical underpinnings are highly various. The unity exists only insofar as that particular grab-bag of cognitive and physical skills has special significance for a community of sea-faring agents. John's `concepts', it seems to me, are just like that: names for complexes of skills whose unity rests not on facts about me, but on facts about John's way of life.

John's tendency to hallucinate his own perspective on to me extends to his conception of my knowledge of the external world. John walks around and feels as if he commands a stable, 3D image of his immediate surroundings. John's feelings notwithstanding, I command no such thing.

I register small regions of detail in rapid succession as I fixate first on this, and then on that aspect of the visual scene. And I do not trouble myself to store all that detail in some internal model in need of constant maintenance and updating. Instead, I am adept at re-visiting parts of the scene so as to re-create detailed knowledge as and when required. As a result of this trick, and others, John has such a fluent capacity to negotiate his local environment that he thinks he commands a constant inner vision of the detail of his surroundings. In truth, what John sees has more to do with the abilities I confer on him to interact constantly, in real time, with rich external sources of information than with the kind of passive and enduring registration of information in terms of which he conceives his own seeings.

The sad fact, then, is that almost nothing about me is the way John imagines it to be. We remain strangers despite our intimacy (or perhaps because of it). John's language, introspections, and over-simplistic physicalism incline him to identify my organization too closely with his own limited perspective. He is thus blind to my fragmentary, opportunistic and generally alien nature. He forgets that I am in large part a survival- oriented device which greatly pre-dates the emergence of linguistic abilities, and that my role in promoting conscious and linguaform cognition is just a recent sideline. This sideline is, of course, a major root of his misconceptions. Possessed as John is of such a magnificent vehicle for the compact and communicable expression of knowledge, he often mistakes the forms and conventions of that vehicle for the structure of thought itself.

But hope springs eternal (more or less). I am of late heartened by the emergence of new investigative techniques such as non-invasive brain imaging, the study of artificial neural networks, and the use of real-world robotics. Such techniques bode well for a better understanding of the very complex relations between my activity, the local environment, and the patchwork construction of the sense of self. In the meantime, just bear in mind that despite our intimacy, John really knows very little about me. Think of me as the Martian in John's head.

Notes

[1] The ideas and themes pursued in this little fantasy owe much to the visions of P.M. and P.S. Churchland, Daniel Dennett, Marvin Minsky, Gilbert Ryle, John Haugeland and Rodney Brooks. In bringing these themes together I have tried for maximum divergence between agent- and brain-level facts. I do not mean to claim dogmatically that current neuroscience unequivocally posits quite such a radical divergence. Several of the issues on which I allow the brain to take a stand remain the subject of open neuroscientific debate. For a taste of the debate, see P.S. Churchland and T.J. Sejnowski, The Computational Brain (Cambridge, MA: MIT Press, 1992) and P.S. Churchland, V.S. Ramachandran and T.J. Sejnowski, `A critique of pure vision', in Large-scale Neuronal Theories of the Brain, ed. C. Koch and J. Davis (Cambridge, MA: MIT Press, 1994).

Explicit supporting references seemed out of place given the literary conceit adopted, but they would include especially: D. Dennett, Brainstorms (Cambridge, MA: MIT Press, 1980), D. Dennett, Consciousness Explained (Boston, MA: Little Brown, 1991), M. Minsky, The Society of Mind (New York: Simon & Schuster, 1985), P.M. Churchland, A Neurocomputational Perspective (Cambridge, MA: MIT Press, 1989), J. Haugeland, `Mind embodied and embedded',

in Mind and Cognition: Proceedings of the First International Conference on Mind and Cognition, ed. Yu-Houng Houng,(Taipei, Taiwan: Academia Sinica, to appear), R. Brooks, `Intelligence without representation', Artificial Intelligence, 41 (1991), pp. 139–59, G. Ryle, The Concept of Mind (London: Hutchinson, 1949) and C. Warrington and R. McCarthy, `Categories of knowledge; further fractionations and an attempted integration', Brain, 110 (1987), pp. 1273–96.

For my own pursuit of some of these themes, see: A. Clark, Associative Engines: Connectionism, Concepts and Representational Change (Cambridge, MA: MIT Press, 1993) and A. Clark, `Moving minds: situating content in the service of real-time success', in Philosophical Perspectives, 10, ed. J. Tomberlin (Atascadero, CA: Ridgeway, forthcoming).

[2] Thanks to Daniel Dennett, Joseph Goguen, Keith Sutherland, David Chalmers and an anonymous referee for support, advice and suggestions.

[3] Or Mary's, or Mariano's, or Pepa's. The choice of the classic male English name is intended only as a gentle reference to those old Readers Digest articles with titles like, `I am John's Liver', `I am John's Kidney', etc.

 jcs-online menu