Conscious Entities

The Archive

2010 to 2012

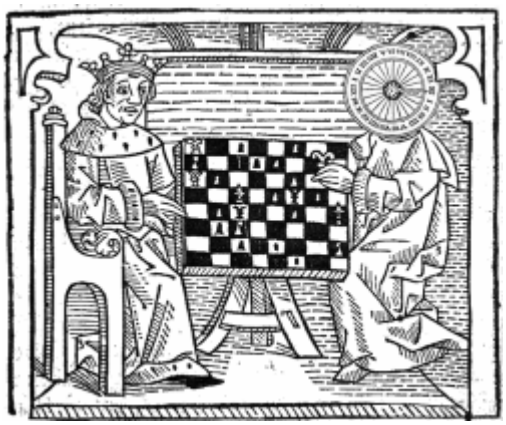
Welcome to Conscious Entities, the Archive

This is the third in a set of documents which bring together the posts which formed ‘Conscious Entities’, my blog about consciousness, which ran from 2004 to 2021(with a final postscript). A few posts have been omitted, and quite a few are rather out of date, but I still enjoy reading them. I’m not really expecting anyone else to be very interested, but I neither wanted to keep the blog frozen indefinitely nor let the content disappear forever, so this is my half-baked solution.

To repeat what I said on the blog: these pieces are devoted to short discussions of some of the major thinkers and theories about consciousness (as they were at the time). The selection of topics is idiosyncratic and the coverage is not systematic. The pieces are meant to be brief and lively, and written from a distinct point of view (sometimes more than one point of view), and this naturally makes it difficult at times to do full justice to the views or the people under consideration. But no-one is mentioned here who I do not sincerely respect and admire, however much I may think they are mistaken on some points. Clearly the only way to get a full understanding of what someone is saying is to read their own books or papers, a course I heartily recommend.

Buy AI

10 January 2010



Kenneth Rogoff is putting his money on AI to be the new source of economic growth, and he seems to think the Turing Test is pretty much there for the taking.

His case is mainly based on an analogy with chess, where he observes that since the landmark victory of "Deep Blue" over Kasparov, things have continued to move on, so that computers now move in a sphere far above their human creators, making moves whose deep strategy is impenetrable to merely human brains. They can even imitate the typical play of particular Grandmasters in a way which reminds Rogoff of the Turing Test. If computers can play chess in a way indistinguishable from that of a human being, it seems they have already passed the 'Chess Turing Test'. In fact he says that nowadays it takes a computer to spot another computer.

I wonder if that's literally the case: I don't know much about chess computing, but I'd be slightly surprised to hear that computer-detecting algorithms as such had been created. I think it's more likely that where a chess player is accused of using illicit computer advice, his accusers are likely to point to a chess program which advises exactly the moves he made in the particular circumstances of the game. Aha, they presumably say, those moves of yours which turned out so well make no sense to us human beings, but look at what the well-known top-notch program Deep Gambule 5000 recommends...

There's a kind of melancholy pleasure for old gits like me in the inversion which has occurred over chess; when we were young, chess used to singled out as a prime example of what computers couldn't do, and the reason was usually given as being the combinatorial explosion which arises when you try to trace out every possible future move in a game of chess. For a while people thought that more subtle programming would get round this, but the truth is that in the end the problem was mainly solved by sheer brute force; chess may be huge, but the computing power of contemporary computers has become even huger.

On the one hand, that suggests that Rogoff is wrong. We didn't solve the chess problem by endowing computers with human-style chess reasoning; we did it by throwing ever bigger chunks of data around at ever greater speeds. A computer playing grandmaster chess may be an awesome spectacle, but not even the most ardent computationalist thinks there's someone in there. The Turing Test, on the other hand, is meant to test whether computers could think in broadly the human way; the task of holding a conversation is supposed to be something that couldn't be done without human-style thought. So if it turns out we can crack the test by brute force (and mustn't that be theoretically possible at some level?) it doesn't mean we've achieved what passing the test was supposed to mean.

In another way, though, the success with chess suggests that Rogoff is right. Some of the major obstacles to human-style thought in computers belong to the family of issues related to the frame problem, in its broadest versions, and the handling of real-world relevance. These could plausibly be described as problems with combinatorial explosion, just like the original chess

issue but on a grander scale. Perhaps, as with chess, it will finally turn out to be just a matter of capacity?

All of this is really a bit beside Rogoff's main interest; he is primarily interested in new technology of a kind which might lead to an economic breakthrough; although he talks about Turing, the probable developments he has in mind don't actually require us to solve the riddle of consciousness. His examples; "from managing the electronics and lighting in our homes to populating "smart grids" for water and electricity, helping monitor these and other systems to reduce waste" actually seem like fairly mild developments of existing techniques, hardly the sort of thing that requires deep AI innovation at all. The funny thing is, I'm not sure we really have all that many really big, really new ideas for what we might do with the awesome new computing power we are steadily acquiring. This must certainly be true of chess - where do we go from here, keep building even better programs to play games against each other, games of a depth and subtlety which we will never be able to appreciate?

There's always the Blue Brain project, of course, and perhaps CYC and similar mega-projects; they can still absorb more capacity than we can yet provide. Perhaps in the end consciousness is the only worthy target for all that computing power after all.

Phi

24 January 2010

I was wondering recently what we could do with all the new computing power which is becoming available. One answer might be calculating phi, effectively a measure of consciousness, which was very kindly drawn to my attention by Christof Koch. Phi is actually a time- and state-dependent measure of integrated information developed by Giulio Tononi in support of the Integrated Information Theory (IIT) of consciousness which he and Koch have championed. Some readable expositions of the theory are here and here with the manifesto here and a formal paper presenting phi here. Koch says the theory is the most exciting conceptual development he's seen in "the inchoate science of consciousness", and I can certainly see why.

The basic premise of the theory is simply that consciousness is constituted by integrated information. It stems from the phenomenological observations that there are vast numbers of possible conscious states, and that each of them appears to unify or integrate a very large number of items of information. What really lifts the theory above the level of most others in this area is the detailed mathematical under-pinning, which means phi is not a vague concept but a clear and possibly even a practically useful indicator.

One implication of the theory is that consciousness lies on a continuum: rather than being an on-or-off matter, it comes in degrees. The idea that lower levels of consciousness may occur when we are half-awake, or in dogs or other animals, is plausible and appealing. Perhaps a little less intuitive is the implication that there must be in theory be higher states of consciousness than any existing human being could ever have attained. I don't think this means states of greater intelligence or enlightenment, necessarily; it's more a matter of being more awake than awake, an idea which (naturally enough, I suppose) is difficult to get one's head around, but has a tantalising appeal.

Equally, the theory implies that some minimal level of consciousness goes a long way down to systems with only a small quantity of integrated information. As Koch points out, this looks like a variety of panpsychism or panexperientialism, though I think the most natural interpretation is that real consciousness probably does not extend all that far beyond observably animate entities.

One congenial aspect of the theory for me is that it puts causal relations at the centre of things: while a system with complex causal interactions may generate a high value of phi, a 'replay' of its surface dynamics would not. This seems to capture in a clearer form the hand-waving intuitive point I was making recently in discussion of Mark Muhlestein's ideas. Unfortunately, calculation of Phi for the human brain remains beyond reach at the moment due to the unmanageable levels of complexity involved; this is disappointing, but in a way it's only what you would expect. Nevertheless, there is, unusually in this field, some hope of empirical corroboration.

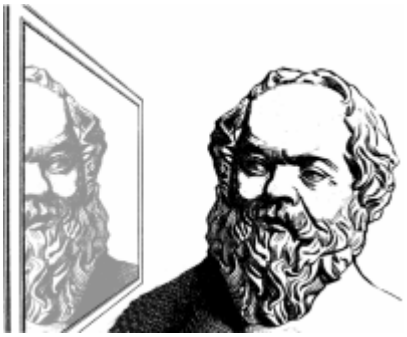
I think I'm convinced that phi measures something interesting and highly relevant to consciousness; perhaps it remains to be finally established that what it measures is consciousness itself, rather than some closely associated phenomenon, some necessary but

not sufficient condition. Your view about this, pending further evidence, may be determined by how far you think phenomenal experience can be identified with information. Is consciousness in the end what information - integrated information - just feels like from the inside? Could this be the final answer to the insoluble question of qualia? The idea doesn't strike me with the 'aha!' feeling of the blinding insight, but (and this is pretty good going in this field) it doesn't seem obviously wrong either. It seems the right kind of answer, the kind that could be correct.

Could it?

Introspection

7 February 2010



Introspection, the direct examination of the contents of our own minds, seems itself to be in many minds at the moment. The latest issue of the Journal of Consciousness Studies was devoted to papers on introspection, marking the tenth anniversary of the publication of The View from Within, by Francisco Varela and Jonathan Shear (which was itself a special edition of the JCS); and now Eric Schwitzgebel has produced a new entry for the Stanford Encyclopaedia of Philosophy.

The two accounts are of course quite different in some respects. The encyclopaedia entry is a careful, scholarly account, neutral and comprehensive; the JCS issue is openly a rallying-cry in support of a programme flowing from Varela's work. This, it seems, called for an end to the ban on examination of lived experience; the JCS gives the impression that it was something of a milestone, though Schwitzgebel's piece does not mention it (he does cite an earlier paper by Varela, once again in the JCS).

What's all this about a ban? Well, back in the nineteenth century, psychologists had no fears about using introspective evidence; it was thought that a proper scientific effort would lead to an objectively verifiable kind of phenomenology. We should be able to classify the elements of mental experience and clarify how they worked together, just by examining what went on in our own heads. A great deal of work was done on all this (It was a great disappointments for me to discover, on first opening Brentano's Psychology from an Empirical Standpoint, that it consisted almost entirely of this kind of thing, and that the only passage about intentional inexistence, the interesting issue, was the couple of paragraphs which I had already read as quotes in several other books.). There was a gradual refinement of the methods involved, leading on to the great heyday of introspectionism, with Wundt and Titchener in the lead. Unfortunately, it became clear that the rival schools of introspectionism had begun to come up with results which in some respects were radically different and incompatible, and since our own introspections are by their nature private and unverifiable, all they could really do by way of settling the issues was to shout at each other.

This embarrassing impasse led to a reaction away from introspection and to the rise of behaviourism, which not only denied the usefulness of examining our inner experience, but actually went to the extreme of denying that there was any such thing as inner experience. Behaviourism in its turn fell out of favour, but according to Varela there remained an instinctive distrust of introspection which continued to put people off it as an avenue of research. This is the 'ban' he wanted to see overturned.

Was there, is there, really a ban? Not exactly. Apart from the most dogmatic of the behaviourists, no-one has ever tried to exclude introspection altogether. In recent times, introspective evidence has been widely accepted - the problem of qualia, thought by some to be the problem of consciousness, depends entirely on introspection. I think the real problem arises when we adopt special methods. In order to obtain consistent results, the old introspectionists thought extensive training was necessary. It wasn't enough to sit and think for a bit; you had to have mastered certain skills of discrimination and perception. The

methodological dangers involved in teaching your researchers what kind of thing they could legitimately look for are clear.

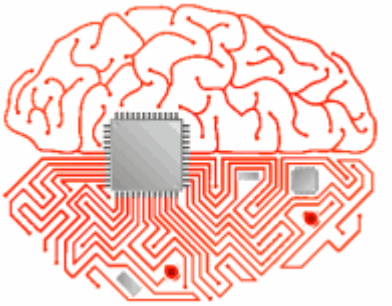
Unfortunately, it seems to be very much this kind of programme which the JCS authors would like to resurrect - or rather, have resurrected, and wish to gain acceptance and support for. Once again we are going to need to learn how to introspect properly before our observations will be acceptable. What makes it worse for me is that the proposal seems to be tied up with NLP - Neuro-linguistic Programming. I don't know a great deal about NLP: it seems to be a protean doctrine which shares with the Holy Roman Empire the property of not really being any of the three things in its name - but for me it does nothing to render another trip down this particular blind alley more attractive.


I don't know about that, but aren't they right to emphasise the potential value of introspection? Isn't it the case that introspection is our only source of infallible information? Most of the things we perceive are subject to error and delusion, but we can't, for example, be wrong about the fact that we are feeling pain, can we? Our impressions of the outside world come to us through a chain of cause and effect, and at any stage errors or misinterpretations can creep in; but because introspection is direct, there's no apce for error to occur. You could well say it's our only source of certain knowledge - isn't that worth pursuing a little more systematically?


That is the exact reverse of the truth: in fact all introspections are false. Think about it. Introspection can only address the contents of consciousness, right? You can't introspect the unconscious mental processes that keep you balanced, or regulate your heartbeat. But all of the contents of consciousness have intentionality - they're about things, yes? So to have direct experience of mental content is to be thinking about something else - not about the mental item itself, but about the thing it's about! Now when the attempt to think directly about our own mental states, it follows that we're necessarily imagining them. Far from having direct contact, we are inevitably thinking about something we've just made up.

AI Resurgent

25 February 2010



Where has AI (or perhaps we should talk about AGI) got to now? h+ magazine reports remarkably buoyant optimism in the AI community about the achievement of Artificial General Intelligence (AGI) at a human level, and even beyond. A survey of opinion at a recent conference apparently showed that most believed that AGI would reach and surpass human levels during the current century, with the largest group picking out the 2020s as the most likely decade. If that doesn't seem optimistic enough, they thought this would occur without any additional fundingfor the field, and some even suggested that additional money would be a negative, distracting factor.

Of course those who have an interest in AI would tend to paint a rosy picture of its future, but the survey just might be a genuine sign of resurgent enthusiasm, a second wind for the field ('second' is perhaps understating matters, but still). At the end of last year, MIT announced a large-scale new project to 're-think AI'. This Mind Machine Project involves some eminent names, including none other than Marvin Minsky himself. Unfortunately (following the viewpoint mentioned above) it has $5 million of funding.

The Project is said to involve going back and fixing some things that got stalled during the earlier history of AI, which seems a bit of an odd way of describing it, as though research programmes that didn't succeed had to go back and relive their earlier phases. I hope it doesn't mean that old hobby-horses are to be brought out and dusted off for one more ride.

The actual details don't suggest anything like that. There are really four separate projects:¶

- Mind: Develop a software model capable of understanding human social contexts- the signpost that establish these contexts, and the behaviors and conventions associated with them. Research areas: hierarchical and reflective common sense. Lead researchers: Marvin Minsky, Patrick Winston.
- Body: Explore candidate physical systems as substrate for embodied intelligence. Research areas: reconfigurable asynchronous logic automata, propagators. Lead researchers: Neil Gershenfeld, Ben Recht, Gerry Sussman.
- Memory: Further the study of data storage and knowledge representation in the brain; generalize the concept of memory for applicability outside embodied local actor context. Research areas: common sense. Lead researcher: Henry Lieberman.
- Brain and Intent: Study the embodiment of intent in neural systems. It incorporates wet laboratory and clinical components, as well as a mathematical modeling and representation component. Develop functional brain and neuron interfacing abilities. Use intent-based models to facilitate representation and exchange of information. Research areas: wet computer, brain language, brain interfaces. Lead researchers: Newton Howard, Sebastian Seung, Ed Boyden

This all looks very interesting. The theory of reconfigurable asynchronous logic automata (RALA) represents a new approach to computation which instead of concealing the underlying physical operations behind high-level abstraction, makes the physical causality apparent: instead of physical units being represented in computer programs only as abstract symbols, RALA is

based on a lattice of cells that asynchronously pass state tokens corresponding to physical resources. I'm not sure I really understand the implications of this - I'm accustomed to thinking that computation is computation whether done by electrons or fingers; but on the face of it there's an interesting comparison with what some have said about consciousness requiring embodiment.

I imagine the work on Brain and Intent is to draw on earlier research into intention awareness. This seems to have been studied most extensively in a military context, but it bears on philosophical intentionality and theory of mind; in principle it seems to relate to some genuinely central and difficult issues. Reading brief details I get the sense of something which might be another blind alley, but is at least another alley.

Both of these projects seem rather new to me, not at all a matter of revisiting old problems from the history of AI, except in the loosest of senses.

In recent times within AI I think there has been a tendency to back off a bit from the issue of consciousness, and spend time instead on lesser but more achievable targets. Although the Mind Machine Project could be seen as superficially conforming with this trend, it seems evident to me that the researchers see their projects as heading towards full human cognition with all that that implies (perhaps robots that run off with your wife?)

Meanwhile in another part of the forest Paul Almond is setting out a pattern-based approach to AI. He's only one man, compared with the might of MIT - but he does have the advantage of not having $5 million to delay his research.

Global Workspace Beats Frame Problem

6 March 2010

Global Workspace theories have been popular ever since Bernard Baars put forward the idea back in the eighties; in 'Applying global workspace theory to the frame problem'*, Murray Shanahan and Baars suggest that among its other virtues, the global workspace provides a convenient solution to that old bugbear, the frame problem.

What is the frame problem, anyway? Initially, it was a problem that arose when early AI programs were attempting simple tasks like moving blocks around. It became clear that when they moved a block, they not only had to update their database to correct the position of the block, they had to update every other piece of information to say it had not been changed. This led to unexpected demands on memory and processing. In the AI world, this problem never seemed too overwhelming, but philosophers got hold of it and gave it a new twist. Fodor, and in a memorable exposition, Dennett, suggested that there was a fundamental problem here. Humans had the ability to pick out what was relevant and ignore everything else, but there didn't seem to be any way of giving computers the same capacity. Dennett's version featured three robots: the first happily pulled a trolley out of a room to save it from a bomb, without noticing that the bomb was on the trolley, and came too; the second attempted to work out all the implications of pulling the trolley out of the room; but there were so many logical implications that it was stuck working through them when the bomb went off. The third was designed to ignore irrelevant implications, but it was still working on the task of identifying all the many irrelevant implications when again the bomb exploded.

Shanahan and Baars explain this background and rightly point out that the original frame problem arose in systems which used formal logic as their only means of drawing conclusions about things, no longer an approach that many people would expect to succeed. They don't really believe that the case for the insolubility of the problem has been convincingly made. What exactly is the nature of the problem, they ask: is it combinatorial explosion? Or is it just that the number of propositions the AI has to sort through to find the relevant one is very large (and by the way, aren't there better ways of finding it than searching every item in order?). Neither of those is really all that frightening; we have techniques to deal with them.

I think Shanahan and Baars, understandably enough, under-rate the task a bit here. The set of sentences we're asking the AI to sort through is not just very large; it's infinite. One of the absurd deductions Dennett assigns to his robots is that the number of revolutions the wheels of trolley will perform in being pulled out of the room is less than the number of walls in the room. This is clearly just one member of a set of valid deductions which goes on forever; the number of revolutions is also less than the number of walls plus one; it's less than the number of walls plus two… It may be obvious that these deductions are uninteresting; but what is the algorithm that tells us so? More fundamentally, the superficial problems are proxies for a deeper concern; that the real world isn't reducible to a set of propositions at all, that, as Borges put it

> *"it is clear that there is no classification of the Universe that is not arbitrary and full of conjectures. The reason for this is very simple: we do not know what thing the universe is."*

There's no encyclopaedia which can contain all possible facts about any situation. You may have good heuristics and terrific search algorithms, but when you're up against an uncategorisable domain of infinite extent, you're surely still going to have problems.

However, the solution proposed by Shanahan and Baars is interesting. Instead of the mind having to search through a large set of sentences, it has a global workspace where things are decided and a series of specialised modules which compete to feed in information (there's an issue here about how radically different inputs from different modules manage to talk to each other: Shanahan and Baars mention a couple of options and then say rather loftily that the details don't matter for their current purposes. It's true that in context we don't need to know exactly what the solution is - but we do need to be left believing that there is one).

Anyway, the idea is that while the global workspace is going about its business each module is looking out for just one thing. When eventually the bomb-is-coming-too module gets stimulated, it begins sending very vigorously and that information gets into the workspace. Instead of having to identify relevant developments, the workspace is automatically fed with them.

That looks good on the face of it; instead of spending time endlessly sorting through propositions, we'll just be alerted when it's necessary. Notice, however, that instead of requiring an indefinitely large amount of time, we now need an indefinitely large number of specialised modules. Moreover, if we really cover all the bases, many of those modules are going to be firing off all the time. So when the bomb-is-coming-too module begins to signal frantically, it will be competing with the number-of-rotations-is-less-than-the-number-of-walls module and all the others and will be drowned out. If we only want to have relevant modules, or only listen to relevant signals, we're back with the original problem of determining just what is relevant.

Still, let's not dismiss the whole thing too glibly. It reminded me to some degree of Edelman's analogy with the immune system, which in a way really does work like that. The immune system cannot know in advance what antibodies it will need to produce, so instead it produces lots of random variations; then when one gets triggered it is quickly reproduced in large numbers. Perhaps we can imagine that if the global workspace were served by modules which were not pre-defined, but arose randomly out of chance neural linkages, it might work something like that. However, the immune system has the advantage of knowing that it has to react against anything foreign, whereas we need relevant responses for relevant stimuli. I don't think we have the answer yet.

Google Consciousness

14 March 2010


I was interested to see a Wired piece recently with points about how Google picks up contextual clues. I've heard before about how Google's translation facilities basically use the huge database of the web: instead of applying grammatical rules or anything like that, they just find equivalents in parallel texts, or alternatives that people use when searching, and this allows them to do a surprisingly good - not perfect - job of picking up those contextual issues that are the bane of most translation software. At least, that's my understanding of how it works. Somehow it hadn't quite occurred to me before, but a similar approach lends itself to the construction of a pretty good kind of chatbot - one that could finally pass the Turing Test unambiguously.


Ah, the oft-promised passing of the Turing Test. Wake me up when it happens - we've been round this course so many times in the past.


Strangely enough, this does remind me of one of the things we used to argue about a lot in the past. You've always wanted to argue that computers couldn't match human performance in certain respects in principle. As a last resort, I tried to get you to admit that in principle we could get a computer to hold a conversation with human-level responses just by the brutest of brute force solutions. You just can a perfect response for every possible sentence. When you get that sentence as input, you send the canned response as output. The longest sentence ever spoken is not infinitely long, and the number of sentences of any finite length is finite; so in principle we can do it.


I remember: what you could never grasp was that the meaning of a sentence depends on the context, so you can't devise a perfect response for every sentence without knowing what conversation it was part of. What would the canned response be to; 'What do you mean?' - to take just one simple example.


What you could never grasp was that in principle we can build in the context, too. Instead of just taking one sentence, we can have a canned response to sets of the last ten sentences if we like - or the last hundred sentences, or whatever it takes. Of course the resources required get absurd, but we're talking about the principle, so we can assume whatever resources we want. The point I wanted to make is that by using the contents of the Internet and search enquiries, Google could implement a real-world brute-force solution of broadly this kind.


I don't think the Internet actually contains every set of a hundred sentences ever spoken during the history of the Universe.

 No, granted; but it's pretty good, and it's growing rapidly, and it's skewed towards the kind of thing that people actually say. I grant you that in practice there will always be unusual contextual clues that the Google chatbot won't pick up, or will mishandle. But don't forget that human beings miss the point sometimes, too. It seems to me a realistic aspiration that the level of errors could fairly quickly be pushed down to human levels based on Internet content.

 It would of course tell us nothing whatever about consciousness or the human mind; it would just be a trick. And a damaging one. If Google could fake human conversation, many people would ascribe consciousness to it, however unjustifiably. You know that quite poor, unsophisticated chatbots have been treated by naive users as serious conversational partners ever since Eliza, the grandmother of them all. The internet connection makes it worse, because a surprising number of people seem to think that the Internet itself might one day accidentally attain consciousness. A mad idea: so all those people working on AI get nowhere, but some piece of kit which is carefully designed to do something quite different just accidentally hits on the solution? It's as though Jethro Tull had been working on his machine and concluded it would never be a practical seed-drill; but then realised he had inadvertently built a viable flying machine. Not going to happen. Thing is, believing some machine is a person when it isn't is not a trivial matter, because you then naturally start to think of people as being no more than machines. It starts to seem natural to close people down when they cease to be useful, and to work them like slaves while they're operative. I'm well aware that a trend in this direction is already established, but a successful chatbot would make things much, much, worse.

 Well, that's a nice exposition of the paranoia which lies behind so many of your attitudes. Look, you can talk to automated answering services as it is: nobody gets het up about it or starts to lose their concept of humanity.

Of course you're right that a Google chatbot in itself is not conscious. But isn't it a good step forward? You know that in the brain there are several areas that deal with speech; Broca's area seems to put coherent sentences together while Wernicke's area provides the right words and sense. People whose Wernicke's area has been destroyed, but who still have a sound Broca's area apparently talk fluently and sort of convincingly, but without ever really making sense in terms of the world around them. I would claim that a working Google chatbot is in essence a Broca's area for a future conscious AI. That's all I'll claim, just for the moment.

Heidegger Vindicated?

27 March 2010

A paper by Dotov, Nie, and Chemero describes experiments which it says have pulled off the remarkable feat of providing empirical, experimental evidence for Heidegger's phenomenology, or part of it; the paper has been taken by some as providing new backing for the Extended Mind theory, notably expounded by Andy Clark in his 2008 book ('Supersizing the Mind').

Relating the research so strongly to Heidegger puts it into a complex historical context. Some of Heidegger's views, particularly those which suggest there can be no theory of everyday life, have been taken up by critics of artificial intelligence. Hubert Dreyfus in particular, has offered a vigorous critique drawing mainly from Heidegger an idea of the limits of computation, one which strongly resembles those which arise from the broadly conceived frame problem, as discussed here recently. The authors of the paper claim this heritage, accepting the Dreyfusard view of Heidegger as an early proto-enemy of GOFAI .

For it is GOFAI (Good Old-Fashioned Artificial Intelligence) we're dealing with. The authors of the current paper point out that the Heideggerian/Dreyfusard critique applies only to AI based on straightforward symbol manipulation (though I think a casual reader of Dreyfus could well be forgiven for going away with the impression that he was a sceptic about all forms of AI), and that it points toward the need to give proper regard to the consequences of embodiment.

Hence their two experiments. These are designed to show objective signs of a state described by Heidegger, known in English as 'ready-to-hand'. This seems a misleading translation, though I can't think of a perfect alternative. If a hammer is 'ready to hand', I think that implies it's laid out on the bench ready for me to pick it up when I want it; the state Heidegger was talking about is the one when you're using the hammer confidently and skilfully without even having to think about it. If something goes wrong with the hammering, you may be forced to start thinking about the hammer again - about exactly how it's going to hit the nail, perhaps about how you're holding it. You can also stop using the hammer altogether and contemplate it as a simple object. But when the hammer is ready-to-hand in the required sense, you naturally speak of your knocking in a few nails as though you were using your bare hands, or more accurately, as if the hammer had become part of you.

Both experiments were based on subjects using a mouse to play a simple game. The idea was that once the subjects had settled, the mouse would become ready-to-hand; then the relationship between mouse movement and cursor movement would be temporarily messed up; this should cause the mouse to become unready-to-hand for a while. Two different techniques were used to detect readiness-to-hand. In the first experiment the movements of the hand and mouse were analysed for signs of 1/f? noise. Apparently earlier research has established that the appearance of 1/f? noise is a sign of a smoothly integrated system. The second experiment used a less sophisticated method; subjects were required to perform a simple counting task at the same time as using the mouse; when their performance at this

second task faltered, it was taken as a sign that attention was being transferred to cope with the onset of unreadiness to hand. Both experiments yielded the expected results. (Regrettably some subjects were lost because of an unexpected problem - they weren't good enough at the simple mouse game to keep it going for the duration of the experiment. Future experimenters should note the need to set up a game which cannot come to a sudden halt.)

I think the first question which comes to mind is: why were the experiments were even necessary? It is a common experience that tools or vehicles become extensions of our personality; in fact it has often been pointed out that even our senses get relocated. If you use a whisk to beat eggs, you sense the consistency of the egg not by monitoring the movement of the whisk against your fingers, but as though you were feeling the egg with the whisk, as though there was a limited kind of sensation transferred into the whisk. Now of course, for any phenomenological observation, there will be some diehards who deny having had any such experience; but my impression is that this sort of thing is widely accepted, enough to feature as a proposition in a discussion without further support. Nevertheless, it's true that it this remains subjective, so it's a fair claim that empirical results are something new.

Second, though, do the results actually prove anything? Phenomenologically, it seems possible to me to think of alternative explanations which fit the bill without invoking readiness-to-hand. Does it seem to the subject that the mouse has become part of them, part of a smoothly integrated entity - or does the mouse just drop out of consciousness altogether? Even if we accept that the presence of 1/f? noise shows that integration has occurred, that doesn't give us readiness-to-hand (or if it does, it seems the result was already achieved by the earlier research).

In the second experiment we've certainly got a transfer of attention - but isn't that only natural? If a task suddenly becomes inexplicably harder, it's not surprising that more attention is devoted to it - surely we can explain that without invoking Heidegger? The authors acknowledge this objection, and if I understand correctly suggest that the two tasks involved were easy enough to rule out problems of excessive cognitive load so that, I suppose, no significant switch of attention would have been necessary if not for the breakdown of readiness-to-hand. I'm not altogether convinced.

I do like the chutzpah involved in an experimental attempt to validate Heidegger, though, and I wouldn't rule out the possibility that bold and ingenious experiments along these lines might tell us something interesting.

Monkey See?

11 April 2010

Can monkeys have blindsight? Sean Allen-Hermanson defends the idea in a recent JCS paper. Blindsight is one of those remarkable phenomena that ought to be a key to understanding conscious perception; but somehow we can never quite manage to agree how the key should be turned. Blindsight, to put it very briefly, is where subjects with certain kinds of brain lesions deny seeing something, but can reliably point to it when prompted to try. It's as though the speaking, self-conscious part of the brain is blind, but some other part, well capable of directing the hand when given a chance, can see as well as ever.

There are a number of ways we might account for blindsight. One of the simplest is to suppose that the visual system is degraded but not destroyed in these cases; the signals from the eye are still getting through, but in some way at reduced power. This reduced power level puts them below the limit required for entry into conscious awareness, but they are still sufficient to bias the subject towards the correct response when they are prompted to guess or have a random try. Another popular theory suggests that the effect arises because there are two separate visual channels, only one of which is knocked out in blindsight. There is a good neurological story which can be told in support of this theory, which weighs strongly in its favour; against it, there have been reports of analogous phenomena in the case of other senses, where it is harder to sustain the idea of physically separate channels. Allen-Hermanson cites claims for touch, smell and hearing (I've wondered in the past whether the celebrated deaf percussionist Evelyn Glennie might be an example of "deafhearing"); and even suggestions that the case of alexithymia, in which things not consciously perceived nevertheless cause anxiety or fear, might be similar. It's possible, of course, that blindsight itself comes in more than one form with more than one kind of cause, and that there is something in both these theories - which unfortunately would make matters all the more difficult to elucidate.

For those of us whose main interest in is consciousness, blindsight holds out the tantalising possibility of an experimental route into the mystery of qualia, of what it is for there to be a way something looks. It's tempting to suppose that what is missing in blindsight patients is indeed phenomenal experience. Like the much-discussed zombies, they receive the data from their senses and are able to act on it but have no actual experience. So if we can work out how blindsight works, we've naturalised qualia and the hard problem is cracked...

Well, no, of course it isn't really that easy. The point about qualia, strictly interpreted, is that they don't cause actions; qualia-free zombies behave just the same as normal people, and that includes speech behaviour. So the absence of qualia could have no effect on what you say; since whatever blindsight patients are missing does affect what they say, it can't be qualia. Moreover we have no conclusive evidence that blindsight patients have no visual experience; it could be that they have the experience but are simply unable to report it. That might seem a strange state to be in, but patients with brain damage are known to assert or confabulate all sorts of things which are at odds with the evidence of their senses; in fact I believe there are subjects who claim with every sign of sincerity to see perfectly when in fact they are demonstrably blind, which is a nice reversal of the blindsight case.

Still, blindsight is a tantalising glimpse of something important about conscious experience, and has all sorts of implications. To pick out one at random, it casts an interesting light on split-brain patients. In blindsight cases, we can have an apparent disconnect between the knowledge the patient expresses with the voice, and the knowledge expressed with the hand; that's pretty much what we get in many experiments on split-brain patients (since normally only one hemisphere has use of the vocal apparatus and the other can only express itself by hand movements). Any claims that split-brain patients are therefore shown to be two different people in a single skull are undercut unless we're willing to take up the unlikely position that blindsight patients are also split people.

One interesting extension of blindsight research is the apparent discovery by Cowey and Stoerig of the same phenomenon in monkeys. There is an obvious difficulty here, since human blindsight experiments typically rely on the subject to report in words what they can see, something monkeys can't do. Cowey and Stoerig devised two experiments; in the first the monkeys were trained to touch a screen where a stimulus appeared; all were able to do this without problems. In the second experiment, the stimulus did not always appear on cue; when it did not, the monkeys were required to press a separate button. Normal monkeys could do this without difficulty, but monkeys with lesions thought to be analogous to those causing blindsight now went wrong when the stimulus appeared in their blind spot, hitting the 'no stimulus' button. Taking the two experiments together, it was concluded that blindsight was effectively demonstrated; the damaged monkeys who could earlier touch the right part of the screen even when the stimulus was in their blind spot, later 'reported' the same stimulus as absent.

(Readers may wonder about the ethical propriety of damaging the brains of living primates for these experiments; I haven't read the original papers, but I suppose we must assume that at any rate the experiments had medical as well as merely philosophical value.)

Of course, these experiments differ significantly from those carried out on human subjects, and as Allen-Hermanson reports, reasonable doubts were subsequently raised in a 2006 paper by Mole and Kelly, who pointed out that relying on two separate experiments, which made differing demands, made the results inconclusive. In particular, the second task was more complex than the first, and it could plausibly be argued that the result of having to deal with this additional complexity was that the monkeys simply failed to notice in the second experiment the stimulus they had picked up successfully in the first.

Allen-Hermanson's aim is to rescue Cowey and Stoerig's conclusions, while acknowledging the validity of the criticisms. He proposes a new experiment: first the monkeys are trained to press a green button if there is a stimulus (no need to point to where it is any more), and a red one if there is none. Then we introduce two different stimuli: Xs and Os. Both the green and red buttons are now divided in two, one side labelled for X, the other for O. If there is a stimulus, the monkeys must now press either green X or green O depending on which appeared: if there is no stimulus, they can press either red button. Allen-Hermanson believes the blindsighted monkeys will consistently press red X correctly if the stimulus is X, even though they are effectively asserting that there is no stimulus.

Maybe. I can't help feeling that all the monkeys will be puzzled by a task which effectively asks them to state whether a stimulus is present, and then, if not present, say whether it was an X or O. The experiment has not been carried out; but Allen-Hermanson goes on to suggest that Mole and Kelly's alternative hypothesis is actually implausible on other grounds. On their interpretation, for example, the blindsighted monkeys simply fail to notice a stimulus in their

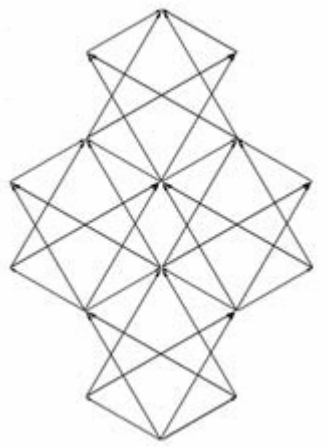
blind spot: yet it has been demonstrated that they cannot recognise objects as salient in monkey terms as ripe fruit when they are presented to the blind spot - so it seems unlikely that we're dealing with something as simple as inattention.

What would it mean if monkeys did have blindsight? It would seem to show, at least, that monkeys are not automata; that they do have something which corresponds to at least one important variety of human consciousness. Allen-Hermanson proposes working further along the mammalian line, and he seems to expect that mammals and even some other vertebrates would yield similar results (he draws the line at toads).

At any rate, we're left feeling that human consciousness is not as unique as it might have seemed. I can't help also feeling more strongly than before that the really unique feature of human awareness is the way it is shot through with language; we may not have the only form of consciousness, but we certainly seem to have the talkiest.

Decommissioning Physicalism

26 April 2010



Panpsychism, or perhaps it would be more accurate to say panexperientialism, has seemed to be quite a popular view in previous discussions here, but I've always found it problematic. Panexperientialism, as the name suggests, is the belief that experience is everywhere; that experience is the basis of reality, out of which everything else is built. Objects which seem dead and inanimate ultimately consist of experience just as much as we do: it just doesn't seem like that because the experiences which make them up are not our experiences. The attractive feature of this view is that it removes some of the mystery from consciousness: instead of being a very rare phenomenon which only occurs in very specific circumstances, such as those which exist in our brains, consciousness of a sort is universal, and so it' s not at all surprising that we ourselves are conscious.

One of the problems is the question of how many experiential loci we're dealing with. Does the table have experiences? Does half the table? Do the table legs have four separate sets of experiences, and at the same time a sort of federal joint experience as a composite entity? There are ways to solve these problems, but they're distinctly off-putting to me. More fundamentally, I'm inclined to doubt whether the theory is as helpful in explaining things as it seems. OK, so my brain has experience just because it's an object and all objects have experience; but surely that brain-as-an-object experience is the same kind of thing as the experiences rocks have, while the experience I'm interested in, the kind that influences my bodily behaviour, is something else; something which remains unexplained.

One of the sources of difficulty here, I think, as with many metaphysical theories, is that the philosophical point of view is not well integrated with any clear scientific conception. When we need to pin down our loci of experience we're left to rummage around and see what we can come up with - atoms? Too small. Discrete physical entities? What exactly are they? (Shintoism, if I understand it correctly, has bitten this kind of bullet and given up on a sharp definition of what is animate: lots of things can have souls, but only if they're salient or impressive. Mount Fuji definitely gets one, but some anonymous pile of dirt in your back garden is just a pile of dirt.)

But what about Finite Eventism? This theory (as expounded by Carey R. Carlson in Chapter 12 of Mind that abides: Panpsychism in the New Millennium) is a theory of physics first and foremost, one that happens to provide a neat basis for a solution to the mind-body problem taking a panpsychist/panexperientialist view. It is based on the late ideas of Russell and Whitehead, though one of its appealing features is that it dovetails well with quantum theory in a way Russell and Whitehead were not aware of.

The gist of the theory is a radical reduction of physics to a minimalist ontology consisting of events and the basic temporal relations of being earlier or later (there's also cause and effect, which I take to be causally connected varieties of earlier/later). There are some rules which prevent inconsistency (events can't be earlier/later than themselves) and that's it. In particular, there is no initial concept of space or extension. We are allowed convergent and divergent causal paths, so we can construct complex multiple 'honeycomb' pathways like the one shown.

The four consistent axes which appear in these diagrams are taken to make up a 4-D manifold of space-time, with neutrinos and electrons delineated by gaps, and the repetition of patterns in sequence representing persistence over time. Quanta, in this theory, are represented by the steps between two events; the number of intermediate steps between two events corresponds with the relative frequency of the relevant path, and these relative frequency ratios provide relative energy ratios, following Planck's E=hf.

I hope those brief, inadequate remarks give a hint of how the basics of physics can be built up in a very elegant manner from the simple topology of these sequences: it looks impressive, though I must frankly admit that I'm not competent to explain the theory properly, never mind evaluate it. Readers may like to look at this short description (pdf).

The question for us is, what are these 'events'? Carlson follows Whitehead in seeing them all as 'occasions of experience'; in some ways they resemble the monads of Leibniz's radically relativist ontology; they are pure phenomenal moments. Carlson argues that the basis of all science is phenomenal experience; historically, in order to account for those experiences better through Newtonian style physics it became necessary to postulate unexperienced abstract entities; and then the phenomenal experience dropped out of the theory leaving us with a world made of fundamentally unaccountable entities.

The best part of it for me is that the theory provides relatively good and clear answers to the problems I mentioned above. It's clear that the loci of experience are situated in the events: I take it that continuity of experience is guaranteed in exactly the same way as physical continuity by repeating patterns (though what the nature of those patterns amounts to is an interesting question). If that's so, then the relationship between the consciousness of my brain-as-object and the consciousness of my brain-as-brain is also helpfully clarified.

How far, though, is the actual nature of my consciousness clarified? The explanation of most of my mental characteristics is deferred upwards to be explained by the working of the brain. That is, no doubt, exactly as it should be: but it leaves me with no particular reason to adopt panpsychism. The one feature which the theory does explain is phenomenal experience (alright, a fairly important feature!); but it really only does so by telling us that things just do have phenomenal experience. Why should we take it that the events are phenomenal in nature - doesn't ontological parsimony suggest they should be featureless blips?

Still, I think this is the most viable and attractive formulation of panpsychism/panexperientialism I've seen.

Sceptical Sceptical Folk Theory Theory Theory

9 May 2010

Daniel Dennett not sceptical enough about qualia? It seems unlikely. Dennett's trenchant view can be summed up in two words: 'What qualia?''. It makes no sense, he would say, for us to talk about ineffable items of direct experience: things which by their definition, we can't talk about. That's not to say we can't talk about our experience of the world: we just need to talk about it in third-person, heterophenomenological terms. Instead of claiming to discuss people's first-person inner experience, we discuss what they report about their first-person inner experience. In fact, if we think about it carefully, we'll realise that's all we could ever do, all we've ever done: really, whatever we may have supposed, all phenomenology is actually heterophenomenology; all discussion is necessarily in third-person, objective, scrutable, effable terms.

Typically Dennett's is a relatively lonely voice ranged against those who would assert that qualia, direct private phenomenal experiences, are knowably, undeniably real, however hard they may be to explain and to reconcile with the objective third-person world described by science. Now Justin Sytsma ('Dennett's Theory of the Folk Theory', JCS Vol 17, no 3-4) interestingly takes a different tack, suggesting that in fact Dennett has conceded too much by accepting that the folk theory, what ordinary people naively believe about their own experience, includes belief in qualia. He quotes Dennett saying:

"there seem to be qualia, because it really does seem as if science has shown us that colors can't be out there, and hence must be in here…"

Reasonably, if somewhat unphilosophically, Sytsma treats what people actually believe as an empirical matter, something we can test; in a sense we could say that this is turning heterophenomenology on itself. It turns out, apparently, that Dennett's assumption is false: in fact people don't regard, say, redness, as an ineffable mental quality, but as a real property of things in the world. Perhaps ordinary people are more sophisticated than Dennett has given them credit for: perhaps less; not sufficiently aware of 'what science has shown us' for it to have had much impact on what they believe.

What are we to make of this? Well, one issue is that there is an inbuilt tension in the entity we're trying to discuss: Sytsma is talking about folk theories: but folk beliefs are really what we have when we have no theories: a folk theory is a kind of contradiction in terms. Julian of Norwich, I think, said that the worst thing about heretics was that they forced honest Christians to determine the truth of theological propositions which pious folk could otherwise have ignored; in a similar way we might argue that philosophical experimenters force their subjects into addressing tricky phenomenological questions which would otherwise never have troubled them.

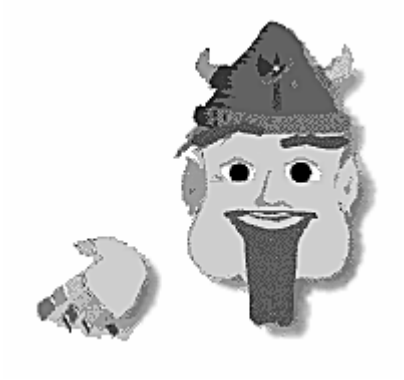
Sytsma, then, by asking his subjects questions, was not evoking their previously held views on phenomenology so much as engendering these views for the first time. There is an obvious danger that the terms of the question would tend to influence the form of the views evoked; but

really that doesn't matter because whatever view Sytsma evoked, it would be different to the no-view that his subjects held initially. Is the folk view pro or anti qualia? The most accurate answer is probably 'no'.

However, forensically Dennett must be in the right. If we want to establish a position, we need to argue against its negation; even if the majority or 'the folk' favour our view, we must instead argue against an arbitrary opponent; in fact against all the most plausible and persuasive opponents we can think of. Even if Dennett was wrong about what people generally believe, tactically it was correct to assume that they disagreed with him: without being unduly elitist, what the majority, or the folk, or the man on the Clapham omnibus actually happen to think, is in this case philosophically uninteresting. We need not worry about whether we should endorse a sceptical theory about Dennett's sceptical theory about folk theories. Isn't that a relief?

Self-Assembly Consciousness

30 May 2010



Kristinn R. Thorisson wants artificial intelligence to build itself. Thorisson was the creator of Gandalf*, the 'communicative humanoid' who was designed in a way that amply disproved Frank Zappa's remark:¶

> "The computer … can give you the exact mathematical design, but what's missing is the eyebrows."¶

Thorisson proposes that constructionism must give way to constructivism (pdf) if significant further progress towards artificial general intelligence is to be made. By constructionism, he means a traditional 'divide and conquer' approach in which the overall challenge is subdivided, modules for specific tasks are more or less hand-coded and then the results are bolted together. This kind of approach, he says, typically results in software whose scope is limited, which suffers from brittleness of performance, and which integrates poorly with other modules. Yet we know that a key feature of general intelligence, and particularly of such features as global attention is a high level of very efficient integration, with different systems sharing heterogeneous data to produce responsive and smoothly coordinated action.

Thorisson considers some attempts to achieve better real-world performance through enhanced integration, including his own, and acknowledges that a lot has been achieved. Moreover, it is possible to extend these approaches further and achieve more: but the underlying problems remain and in some cases get worse: a large amount of work goes into producing systems which may perform impressively but lack flexibility and the capacity for 'cognitive growth'. At best, further pursuit of this line is likely to produce improvements on a linear scale and "Even if we keep at it for centuries… basic limitations are likely to asymptotically bring us to a grinding halt in the not-too-distant future."

It follows that a new approach is needed and he proposes that it will be based on self-generated code and self-organising architectures. Thorisson calls this 'constructivism', which is perhaps not an ideal choice of name, since there are several different constructivisms in different fields already. He does not provide a detailed recipe for constructivist projects, but mentions a number of features he thinks are likely to be important. The first, interestingly, is temporal grounding - he remarks that in contrast to computational systems, time appears to be integral to the operation of all examples of natural intelligence. The second is feedback loops (but aren't they a basic feature of every AI system?); then we have Pan-Architectural Pattern Matching, Small White-Box Components (White-Box as opposed to Black-Box, ie simple modules whose function is not hidden), and Architecture Meta-Programming and Integration.

Whether or not he's exactly right about the way forward, Kristinsson's criticisms of traditional approaches seem persuasive, the more so as he has been an exponent of them himself. They also raise some deeper questions which, as a practical man, he is not concerned with. One issue, indeed, is whether we're dealing here with difficulties in practice or difficulties in principle. Is it just that building a big AGI is extremely complex, and hence in practice just beyond the scope of the resources we can reasonably expect to deploy on a traditional basis?

Or is it that there is some principled problem which means that an AGI can never be built by putting together pre-designed modules?

On the face of it, it seems plausible that the problem is one of practice rather than principle, and is simply a matter of the huge complexity of the task. After all, we know that the human brain, the only example we have of successful general intelligence, is immensely complex, and that it has quirky connections between different areas. This is one occasion when Nature seems to have been indifferent to the principles of good, legible design; but perhaps 'spaghetti code' and a fuzzy allocation of functions is the only way this particular job can be done; if so, it's only to be expected that the sheer complexity of the design is going to defeat any direct attempt to build something similar.

Or we could look at it this way. Suppose constructivism succeeds, and builds a satisfactory AGI. Then we can see that in principle it was perfectly possible to build that particular AGI by hand, if only we'd been able to work out the details. Working out the details may have proved to be way beyond us, but there the thing is: there's no magic that says it couldn't have been put together by other methods.

Or is there? Could it be that there is something about the internal working of an AGI which requires a particular dynamic balance, or an interlocking state of several modules, that can't be set up directly but only approached through a particular construction sequence - one that amounts to it growing itself? Is there after all a problem in principle?

I must admit I can't see any particular reason for thinking that's the way things are, except that if it were so, it offers an attractive naturalistic explanation of how human consciousness might be, as it were, gratuitous: not attributable to any prior design or program, and hence in one sense the furthest back we can push explanation of human thoughts and actions. If that's true, it in turn provides a justification for our everyday assumption that we have agency and a form of free will. I can't help finding that attractive; perhaps if the constructivist approaches Thorisson has in mind are successful this will become clearer in the next few years.

* For anyone worried about the helmet, I should explain that this Gandalf was based on a dwarf from Icelandic cosmogony, not Tolkien's wizard of the same name.

Unconscious Free Will

12 September 2010

Does the idea of unconscious free will even make sense? Paula Droege, in the recent JCS, seems to think it might. Generally experiments like Libet's famous ones, which seemed to show that decisions are made well before the decider is consciously aware of them, are considered fatal to free will. If the conscious activity came along only after the matter had been settled, it must surely have been powerless to affect it (there are some significant qualifications to this: Libet himself, for example, considered there was a power of last-minute veto which he called 'free won't' – but still the general point is clear). If our conscious thoughts were irrelevant, it seems we didn't have any say in the matter.

However, this view implies a narrow conception of the self in which unconscious processes are not really part of me and I only really consist of that entity that does all the talking. Yet in other contexts, notably in psychoanalysis, don't we take the un- or sub-conscious to be more essential to our personality than the fleeting surface of consciousness, to represent more accurately what we 'really' want and feel? Droege, while conceding that if we take the narrow view there's a good deal in the sceptical position, would prefer a wider view in which unconscious acts are valid examples of agency too. She would go further and bring in social influences (though it's not entirely clear to me how the effects of social pressure can properly be transmuted into examples of my own free will), and she offers the conscious mind the consolation prize of being able to influence habits and predispositions which may in turn have a real causal influence on our actions.

I suppose there are several ways in which we exercise our agency. We perhaps tend to think of cases of conscious premeditation because they are the clearest, but in everyday life we just do stuff most of the time without thinking about it much, or very explicitly. Many of the details of our behaviour are left to 'autopilot', but in the great majority of cases the conscious mind would nevertheless claim these acts as its own. Did you stop at the traffic light and then move off again when it turned green? You don't really remember doing it, but are generally ready to agree that you did. In unusual cases, we know that people sometimes even elaborate or confabulate spurious rationales for actions they didn't really determine.

But it's much more muddled than that. We do also at times seek to disown moral responsibility for something done when we weren't paying proper attention, or where our rational responses were overwhelmed by a sudden torrent of emotion. Should someone who responds to the sight of a hated enemy by swerving to collide with the provoker be held responsible because the murderous act stems from emotions which are just as valid as cold calculation? Perhaps, but sometimes the opposite is taken to be the case, and the overwhelming emotion of a *crime passionnel* can be taken as an excuse. Then again few would accept the plea of the driver who says he shouldn't be held responsible for an accident because he was too drunk to drive properly.

I think there may be an analogy with the responsibility held by the head of a corporation: the general rule is that the buck stops with the chief, even if the chief did not give orders for the particular action which subordinates have taken; in the same way we're presumed by default to

be responsible for what we do: but there are cases where control is genuinely and unavoidably lost, no matter what prudent precautions the chief may have put in place. There may be cases where the chief properly has full and sole responsibility; in other cases where the corporation has blundered on in pursuit of its own built-in inclinations it may be appropriate for the organization as a whole to accept blame for its corporate personality: and where confusion reigned for reasons beyond reasonable control, no responsibility may be assigned at all.

If that's right, then Droege is on to something; but if there are two distinct grades of responsibility in play, there ought really to be two varieties of free will; the one exercised explicitly by the fully conscious me, and the other by 'whole person' me, in which the role of the conscious me, while perhaps not non-existent is small and perhaps mostly indirect. This is an odd thought, but if, like Droege, we broadly accept that Libet has disproved the existence of the first variety of free will, it means we don't have the kind we can't help believing in, but do have another kind we never previously thought of - which seems even odder.

Higgs Consciousness

30 June 2010

Is CERN, with its Large Hadron Collider, on the brink of revealing the ultimate nature of consciousness? Deepak Chopra seems to think that the Higgs boson, which CERN's experiments might discover, has something to do with it. 'Could this in fact be where materialism destroys itself from within?' he asks, which seems like one of those Interesting Historical Questions To Which The Answer Is No.

My understanding of the Higgs boson is slight, but I don't really see how Chopra's conclusion follows here: if the existence of the Higgs boson is demonstrated, so far as I know that enables a resolution and tidying-up of some issues in physics, mainly to do with where particles get their mass from. (In my naivety I think I might have been inclined to think that mass was a metaphysical necessity on the grounds that you couldn't make a coherent universe out of things which were under no obligation to keep still…) That seems to be something that can only tend to strengthen materialism; but even if the boson isn't found, we'll merely be looking for a different account of mass. We won't really be any worse off than we are already; arguably better off in having eliminated another blind alley; so if anything materialism is likely to be looking stronger whatever CERN may come up with.

It looks as though Chopra is really just using the Higgs as a pretext for raising again two earlier arguments. First, he thinks mere physics cannot account for the complex structures and the intentionality found in the world; we therefore need to invoke God – not in the form of an old man with a white beard, but a sort of universal consciousness.

I suppose you could connect universal consciousness with Higgs in a way. According to the Higgs theory, mass arises out of a universal Higgs field with certain interesting properties. We could hypothesise the existence of a comparable consciousness field, which endows certain entities passing through it with experience in broadly the way the Higgs field gives particles their mass. It's not a theory I feel inclined to take up, but it might offer a way of developing an attractively clear panexperientialism, one where we might even be able to do some sums and make some predictions about the free-floating momentary experiences which I assume would be the equivalent, in this theory, of the famous boson.

But in fact that's not quite what Chopra means: his second argument is rather that the role of the observer in collapsing wave functions in certain interpretations of quantum physics points us towards a reality in which consciousness is fundamental. He sees this as a step along the road to the perspective of Vedanta, where Brahman the all-inclusive consciousness is the self-interacting dynamic of observer, observed and process of observation.

Even on the most helpful interpretation of quantum physics, I think this reads much too much into the science: in fact I'd go so far as to say that if your reading of quantum physics requires the adoption of universal consciousness, a rather large ontological price, it seems a clear sign that your reading has gone wrong somewhere.

Generally I'd say the march of science is taking us away from universal consciousness rather than towards it. There was a time when it was reasonable to assume that consciousness was naturally out there in divine or panpsychic form and that human consciousness was best explained as some of that natural awareness taking possession of a body. Human consciousness was an inexplicable mystery and referring it to a hypothetical basic element of reality was an economical hypothesis.

Now, however, although consciousness remains mysterious, we can sort of see at least the broad outlines of the sort of way in which it might be naturalised as part of the functioning of a brain. There are some nasty gaps to say the least of it (the 'meaning' which Chopra alludes to can fairly be considered one), and many would still say the job is impossible, or impossible to our limited minds: but we've got a better hypothesis to work on and quite a bit of evidence that we're broadly going the right way.

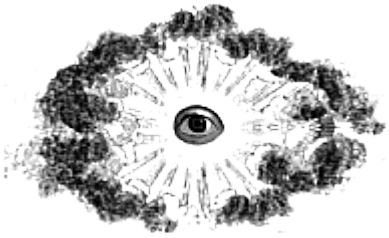
That partial clarification is enough to put the boot on the other foot, and leave us needing more explanation of God, or the universal consciousness. The story of human consciousness as we understand it now relies on the detailed physical interactions of biological material which is itself the product of a process of evolution: but this won't do for divine or universal minds which have no physical structure and never competed for survival: they need some other explanation, of which there is no sign.

This need not be fatal for spiritual or panpsychic explanations: it might offer a bracing challenge and a spur to investigation and new theories. But it certainly suggests to me that waiting for quantum theory to vindicate Vedantic philosophy is, to put it mildly, optimistic.

(PS - The picture is meant to be Higgs, not Chopra.)

The Consciousness Substance

18 July 2010



Mostyn W Jones might seem to be leading with his chin a little when he offers us, in the JCS, "a clear, simple mind-body solution". His "clear physicalism" is meant to banish many of the "obscurities" (a favourite word) involved in other accounts, especially reductionist and functionalist ones.

One problem with functionalism, according to Jones, is that it requires us to believe that conscious sensations are multiply realisable. So long as it embodies the right functions, any physical thing can have consciousness. But functions are abstract; what have they got to do either with my physical brain or with my vivid actual experience? These relations are surely "obscure". Another problem is that none of these reductive theories can deal with qualia (the inherent phenomenal redness of red, hotness of heat, etc). Once you've reduced consciousness to computation, to non-computational functions, or to anything similar, you can no longer explain why it is that real inward experience occurs, or what causal relations that experience has with anything, or how it contrives to have them. Jones sympathises with Strawson and Stoljar: he would like to be able to say that qualia are just the reality of experience: science gives an outside account, and qualia are how it looks from the inside.

In some ways this is an appealing position to take, but there are a number of pitfalls. If all you're saying is that science is the third-person perspective and qualia are the first-person perspective, you're just re-stating the problem: why is there a first-person perspective, anyway? And if qualia are just the reality, you have two options. One is to find an explanation of why they don't crop up all over the place, but seem only to arise in the presence of appropriate brain functions; the other is to bite the bullet and say that in fact they do crop up all over the place, and that panexperientialism is correct. Strawson, if I've understood correctly, leans in this latter direction; Jones, in spite of his hostility to functionalism, seems to lean back the other way: he describes consciousness as a neural substance arising from highly active, highly connected neural circuits; so there still seems to be a broadly functional explanation of why these phenomena are confined to brains.

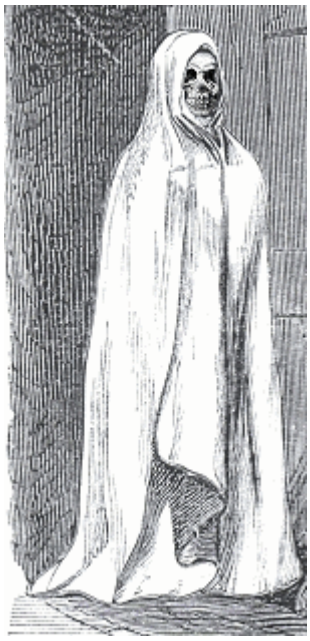
"Substance" is a treacherous word: in ordinary parlance it suggests a lump of stuff, simple physical matter: but in philosophy, especially older philosophy, it has a much more slippery meaning as a basic element, one of the things that remains when analysis is complete. By pursuing this idea of a substance as something unanalysable to its logical conclusion, Leibniz produced the bizarre relativistic ontology of the Monadology: Jones certainly doesn't mean monads, but what does he mean? He says "Clear physicalism avoids this obscurity by treating qualia as electrochemical substances that underlie observable brain activity and do work in brains". The combination of both underlying observable physics yet doing work - causal work, we assume - seems very problematic. It would be odd but compatible with physics if Jones were merely saying that qualia are aspects of the physical world that run along with - perhaps over-determine - the causal effects specified by physics: but that doesn't seem to be quite it.

If we take a step back, the idea of making consciousness a physical substance of any kind seems difficult to accept; consciousness, after all, comes and goes, but matter is conserved. If we take a more flexible view of the notion of substance and allow the substance of consciousness to come and go in line with certain vigorous firings of interlinked neurons it

becomes a little hard to see what our quarrel with some sort of functionalism can be. Pain, it turns out, is not the firing of C-fibres (or whatever), but it is a substance that occurs in perfect correlation with the firing of C-fibres. Hmm.

Qualia: The Movie

24 July 2010



I suppose the zombies couldn't have the film industry all to themselves for ever, and here it is: Qualia, the movie.I wondered at first whether this was something to do with Sony and their Qualia man Ken Mogi, but in fact it seems it is a small independent venture. I said 'here it is', but actually all we have for the moment is a trailer: it seems that the funds required (amounts which I expect wouldn't cover one day's catering budget on a Hollywood blockbuster) have been difficult to get together.

How do you make a film about qualia? (Ken Mogi would probably ask how you could make one without them.) I can't quite decide whether getting ineffable qualities into a film is an amusingly quixotic endeavour or an admirable ambition. It seems all too likely that you would end up with either the talkiest, chin-strokingest film ever made; or an exciting dramatisation of the life of Mary the Colour Scientist. (Susan Blackmore suggested that students should act out this famous gedankenexperiment, after all, though how that would help still rather eludes me.) Actually there's no reason why a film can't at least raise genuine philosophical issues. I'll always remember the Captain's advice on how to deal with the malfunctioning bomb in Dark Star ("Teach it phenomenology"), and The Matrix is often credited with asking interesting questions - though sadly the red and blue pills were soon put aside so that the film could become a kung fu movie performed by people dressed as a Eurythmics tribute band (The Revenge Tourists?).

I haven't found much information about the actual plot of Qualia, but it seems it has something to do with research which triggers or examines ghostly occurrences and disturbs someone's complacent monist materialism. Nothing wrong with disturbing our dogmatic slumbers, of course. I like to think that at some stage a grave scientist will say "Sir! We're detecting... phenomena."

But the association with ghosts is not particularly welcome. I wonder whether this is another sign, like the use of qualia to buttress the theist case, that the hard problem potentially appeals to those who would like the world to be less scientific and more magical. I hope not: I'd hate to see New Age shops selling qualia-enhancing crystals. Perhaps that's just snobbery? After all It's legitimate to claim qualia as evidence for some kind of dualism, and some kind of dualism is what you might well be looking for if you wanted to provide ghosts with some respectable ontological underpinnings. Still, I can only look forward to the film with qualified enthusiasm .

(The film was released in 2012, and I have not heard anything of it since.)

Old Skool Consciousness

1 August 2010

There's an illuminating new piece in the SEP about 17th century theories of consciousness. Your first reaction might be 'what 17th century theories of consciousness?'; the discussion in those days was framed rather differently and it typically requires a degree of interpretation to work out what philosophers of the period actually thought about 'consciousness'. In fact, according to Larry M. Jorgensen, who wrote the entry, the 17th century saw the first emergence of the concept of consciousness as distinct from conscience: in many languages the same word is still used for both.

Hobbes apparently sets this out quite explicitly (somehow this interesting bit must have passed me by when I read Leviathan because it left no impression on my memory); he has conscience originally referring to something which two people knew about ('knew together'), and then metaphorically for the knowledge of one's own secret facts and secret thoughts. Jorgensen tells us that the Cambridge Platonists had a role in developing the modern usage in English where 'conscience' refers to knowledge of one's own moral nature while 'consciousness' means simply knowledge of one's own mental content.

That idea, of having knowledge of one's own mental content, seems to have a reflexive element - we know about what we know; and this was an issue for philosophers of the period, notably Descartes. For Descartes it was essential that my having a thought involved me knowing that I had a thought; but for some this seemed to suggest a second-order theory in which a thought becomes conscious only when accompanied by another thought about the first. Descartes could not accept this: for one thing if knowledge of my own thoughts is not direct, the cogito, Descartes' most famous argument is threatened. The cogito claims that I cannot possibly be wrong about the fact that I am thinking, but if the knowledge of my thought is separate from the thought itself this no longer seems unassailably true.

It seems that while Descartes accepted that awareness of our own thought required some sort of reflection, he denied that the reflection was separate from the thought. He said that [T]he initial thought by means of which we become aware of something does not differ from the second thought by means of which we become aware that we were aware of it.

This can't help but seem a little like cheating - sneaking in an extra thought for nothing. I think the best way to imagine it might be through analogy with a searchlight. We can swing the light around, illuminating here a building, there a tree, just as we can direct our conscious awareness towards different objects. Then Descartes might ask: do we need a second light in order to see the first light? No, of course not, because the light is already illuminated; if the light lights up other objects it must itself be illuminated (if perhaps in not quite the same way).

A surprising amount of Jorgensen's exposition seems to be relevant to current discussions, and not solely because he is, necessarily, reinterpreting it in terms of modern concerns. In some ways I'm afraid we haven't moved on all that much.

Headless Consciousness

6 August 2010

There have been a number of attempts recently to externalise consciousness, or at least extend it beyond the skull. In Out of Our Heads, Alva Noë launches a very broad-based attack on the idea that it's all about the brain, drawing in a wide range of interesting research - mostly relatively well-known stuff, but expounded in a style that is clear and very readable. Unfortunately I don't find the arguments at all convincing; I'm not unsympathetic to extended-mind ideas, but Noë's clear and thorough treatment tended if anything to remind me of reasons why the assumption that consciousness happens in the brain looks so attractive.

I'm happy to go along with Noë on some points: in his first chapter he launches a bit of a side-swipe at scanning technology, fMRI and PET, pugnaciously asking whether it is 'the new phrenology?' and deriding its limitations: this seems a useful corrective to me. But in chapter two we are brought up short by the assertion that bacteria are agents. They have interests, and pursue them, says Noë; they're not just bags of chemicals responding to the presence of sugar. Within their limits we ought to accord them some sort of agency.

To me, proper agency requires an awareness of what acts one is performing and the idea that bacteria could have it at any level seems absurd. How did we get here? Noë's case seems to be that the problem of other minds is effectively insoluble on rational empirical grounds; we can never have really solid reasons for believing anyone else, or any other entity, is conscious; yet we find ourselves unable to entertain seriously the idea that our fellow-humans might be zombies. This, he thinks, is because we have a kind of built-in engagement, almost a kind of moral commitment. He wants to extend this to life fairly widely, and of course if brainless bacteria can have agency, it tends to show that the brain is a bit over-rated. I think he's unnecessarily pessimistic about the evidence for other conscious minds; as a matter of fact books like his are pretty spectacular evidence; how and why would human beings produce such volumes, examining the inner workings of consciousness in minute detail, if they didn't have it? But bacteria have yet to produce such evidence in their own favour.

Noë rests a fair amount of weight on experiments which show the remarkable plasticity of the brain: notably he quotes experiments on ferrets by Mriganka Sur. New-born ferrets had their brains rewired in such a way that the eyes fed into auditory, rather than visual cortex; yet they grew up able to use their eyes perfectly well. This shows, Noë suggests, that no particular part of the brain is required for vision. That might be so, but that in itself does not show that no brain at all will do the job, and obviously it won't: if the ferrets' optical nerves had been linked with their teeth, or left dangling unattached, they would surely have been unambiguously blind. The belief that consciousness is sustained by the brain does not commit us to the view that only one specific set of neurons can do the job. Noë explains, quoting an experiment with a rubber hand, how our sense of our selves and where we are can be moved around in a remarkably vivid way. For him, this shows that where the brain is doesn't matter; but for others it seems equally likely to suggest that what the brain thinks is crucial and where our real hand is doesn't matter at all.

Noë wants to claim that the frequently quoted thought-experiment of a brain in a vat, extracted from the body but still living and thinking, is impossible in principle (we know it's impossible in practice, at least currently). He suggests that even if we did manage to sustain a brain artificially, the supporting vat, providing it with oxygenated blood and all the other complex kinds of support it would need, would actually amount to a new body. This nifty bit of redefinition is meant to show that the idea of a brain without a body will not fly. But the real point here is surely missed. OK, let's accept that the vat is a new body: that still means we can swap bodies while maintaining an individual consciousness. But if we keep the body and swap brains… it just seems impossible to believe that the consciousness wouldn't go with the brain.

This perception seems to me to be the unshiftable bedrock of the discussion. Noë expounds effectively the case for regarding tools and even parts of the environment as parts of our mental apparatus; and he brings in Putnam's argument that 'meaning isn't in the head'. But these arguments only serve to expand our conception of the brain-based mind, not undermine it.

My sympathy for Noë's case returned to some degree when he discussed language. He notes that for Chomsky and others language seemed a miraculous accomplishment because they misconstrued it as an exercise in the formal decoding of a vast array of symbols. In fact, language is an activity rather than a purely intellectual exercise, and develops in a context, pragmatically. I'd go along with that to a great extent, but while Noë sees it as proving that our decoding brain isn't the crux of the matter after all, it seems to me it proves that decoding isn't really what our brains are doing when they process (a word Noë would object to) language.

I sympathised even more with Noë when he attacked the idea that reality is, essentially, an illusion. If this were the case, the brain would be the all-powerful arbiter of reality (although it might seem that if the world is an illusion, the brain must be one too, and we should be dealing with a mind whose actual nature need not be pinkish biological glop). But he seemed to be back on weak ground when he concluded by taking on dreams. Dreams, after all, seem like the perfect evidence that the brain can produce conscious experience without calling on the senses or the body. Noë argues that dreams are more limited than we think, that not all waking experiences can be reproduced in dreams, which are always shifting and inconstant. This might be true, but so what? If the brain can produce conscious experiences on its own - any conscious experiences - that seems to show that, with all caveats duly entered, the brain is still where it really happens.

It's a well-written book, and for someone new to consciousness it would provide many excellent short sketches of thought-provoking experiments and arguments. But I'm staying in my head.

Footnote Consciousness

29 August 2010

I see that the Internet has been getting a certain amount of stick over the way it allegedly alters our mental processes for the worse. Some of this dialogue apparently stems from a two-year-old piece by Nicholas Carr, now developed into a book. Most of the criticisms seem to be from people who have experienced two main problems: they're finding that they have a reduced attention span, and they're also suffering from a failing memory. They attribute these problems to Internet use - but I wonder whether they have made sufficient allowance for the fact that both can also be the result of simple ageing.

I think it's true that if you don't use your memory, it gets worse, so it's superficially plausible that relying on the Internet could have a bad effect: but I don't think I find myself using the Internet for things I would otherwise have learnt by heart, while I certainly have begun forgetting things I knew quite well before the Internet was invented. So far as my attention span is concerned, it has certainly waned steadily over the whole course of my life: when I was four or five I could spend a long time just examining the patterns made by the grain in a piece of wood (mind you, in those days, we had interesting wood, not like the bland stuff they produce these days...). I regret this to some extent, but in another way I don't regret it at all, because I think it is partly a result of mental improvement, the result of accumulated experience. I can tell more quickly now when something is not going to be worth pursuing, and I am less bothered about dropping it quickly. Nowadays, I don't feel at all guilty about dropping a book after chapter one if reading it looks like a mistake, whereas twenty years ago I would no more have stopped reading a book once started than I would have got up and left a dinner party half-way through. I know now that life is too short.

But there are deeper criticisms of the malign effects of the Internet. Jaron Lanier, in an NYT piece (he too has written a book about it; it's interesting that both he and Carr, in spite of the alleged waning of attention spans, still thought this quixotic 'book' business was still worthwhile, rather than just tweeting their thoughts), suggests that we are increasingly deferring to computers, encouraged by inflated claims made for various pieces of software. This has a serious moral dimension - if we see computers as people, we may be led to see people as mere machines - but it also undermines original creativity, a point developed more in the book (OK, I skimmed a few summaries). We start to value mashups and compilations as more valuable than new work generated from scratch. Perhaps worst of all, we may end up letting stupid algorithms make actual decisions for us.

The first of these points is one that has been made before, and I believe it underlies many people's aversion to the whole idea of AI. I think it's undeniable that software producers are gravely inclined to overstate what their programs do, speaking of relatively simple data manipulation as though it involved genuine understanding and originality. But I don't think that has really devalued our conception of humanity - not yet, anyway. Unless and until someone produces a machine which they claim is a conscious being, that remains a danger rather than a current problem. I don't think we're really in danger of delegating important decisions either; letting a computer suggest a track or a book is akin to random browsing of shelves; Lanier

himself notes that even the advocates of the computers don't allow the machines to design their products or run their companies.

There's certainly something in the point about creativity. Hypertext encourages quotation, and I suspect that this has had an influence: an apposite quote is a frequent and respected way of contributing to discussions on popular forums and blogs, to an extent that would seem almost donnish if the quotes weren't typically from Star Wars or the Simpsons rather than Shakespeare. It must surely be the case that sometimes on the Web people use text quotes, photoshopped images and so on when otherwise they would have chosen their own words or drawn their own pictures; but mostly the copied stuff is surely extra. It's a bit like photography; when people could take photographs, they made fewer engravings and oil paintings, but mainly they made many more pictures (and let's be honest - some of those uncreated paintings and engravings were no loss).

There are deeper issues still: has the Web influenced the way we perceive the world? I strongly suspect that films have to some degree influenced the way I see the world and represent my own life to myself. I can't be the only person who has sometimes felt an irresistible urge to do a reaction shot for the benefit of a non-existent audience (one day it may exist if CCTV continues to spread). In one way I think the Internet may have a more pervasive effect. I remarked that the Internet is quotation-driven: but it doesn't just quote, it comments. You could say, I think, that the essence of Web culture is to display something (text, picture, video) and provide comment in parallel (I suppose I'm exemplifying this as I describe it). I suspect that as time goes on reality will come to seem to us like the thing presented and our thoughts like the comments. Our consciousness may end up seeming like a set of lengthy footnotes. Perhaps David Foster Wallace was way ahead of us.

Unconscious Free Will

12 September 2010

An interesting piece by Neil Levy from a few months back, on Does Consciousness Matter? (I only came across it because of its nomination for a 3QD prize). Actually the topic is a little narrower than the title suggests; it asks whether consciousness is essential for free will and moral responsibility (it being assumed that the two are different facets of the same thing - I'm comfortable with that, but some might challenge it). Neil notes that people typically suppose Libet's findings - which suggest our decisions are made before we are aware of them - make free will impossible.

Neil is not actually all that worried by Libet: the impulses from the event of intention formation ought to be registered later than the event itself in any case, he says; so Libet's results are exactly what we should have expected. Again, I'm inclined to agree: making a conscious decision is one thing; the second-order business of being conscious of that conscious decision naturally follows slightly later. (Some, of course, would take a very different view here too; some indeed would say that the second-order awareness is actually what constitutes consciousness.)

Neil particularly addresses two arguments. One says that consciousness is important because only those objectives that are consciously entertained reflect the quality of my will; if I'm not aware that I'm hitting you, I can't be morally responsible for the blows. Neil feels this is a question-begging response which just assumes that conscious awareness is essential; I think he's perhaps a bit over-critical, but of course if we can get a more fully-worked answer, so much the better.

He prefers a slightly different argument which says that factors we were not conscious of cannot influence our deliberations about some act, and hence we can only be held responsible for acts consciously chosen. George Sher, it seems, rejected this argument on the grounds that our actions are influenced by unconscious factors; but Neil rejects this, saying that although unconscious factors certainly influence out behaviour, we have no opportunity to consider them, which is the critical point.

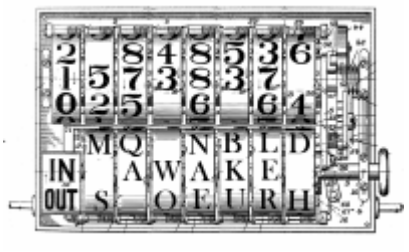
Personally, I would say that agency is inherently a conscious matter because it requires intentionality. In order to form an intention we have to hold in mind (in some sense) an objective, and that requires intentionality. Original intentionality is unique to consciousness - in fact you could argue that it's constitutive of consciousness if you believe all consciousness is consciousness of something - though I myself I wouldn't go quite so far.

But what about those unconscious factors? Subconscious factors would seem to possess intentionality as well as conscious ones, if Freud is to be believed: I don't see how I could hold a subconscious desire to kill my father and marry my mother without those desires being about my parents. Neil would argue that this isn't relevant because we can't endorse, question or reject motives outside consciousness - but how does he know that? If there's unconscious endorsement, questioning and so on going on, he wouldn't know, would he? It could be that the unconscious plays Hyde to the Jekyll of conscious thought, with plans and projects that are in its own terms no less complex or rational than conscious ones.

I think Neil is right: but it doesn't seem securely proved in the way he was hoping for. The unconscious part of our minds never has chance to set down an explanation of its behaviour, after all: it could still in principle be the case that conscious rationalising is privileged in the literature and in ordinary discourse mainly because it's the conscious part of the brain that does the talking and writes the blogs.

The Consciousness Meter

26 September 2010



It has been reported in various places recently that Giulio Tononi is developing a consciousness meter. I think this all stems from a New York Times article by the excellent Carl Zimmer where, to be tediously accurate, Tononi said "The theory has to be developed a bit more before I worry about what's the best consciousness meter you could develop." Wired discussed the ethical implications of such a meter, suggesting it could be problematic for those who espouse euthanasia but reject abortion.

I think a casual reader could be forgiven for dismissing this talk of a consciousness meter. Over the last few years there have been regular reports of scientific mind-reading: usually what it amounts to is that the subject has been asked to think of x while undergoing a scan; then having recorded the characteristic pattern of activity the researchers have been able to spot from scans with passable accuracy the cases where the subject is thinking of x rather than y or z. In all cases the ability to spot thoughts about x are confined to a single individual on a single occasion, with no suggestion that the researchers could identify thoughts of x in anyone else, or even in the same individual a day later. This is still a notable achievement; it resembles (can't remember who originally said this) working out what's going on in town by observing the pattern of lights from an orbiting spaceship; but it falls a long way short of mind-reading.

But in Tononi's case we're dealing with something far more sophisticated. We discussed a few months ago Tononi's Integrated Information Theory (IIT), which holds that consciousness is a graduated phenomenon which corresponds to Phi: the quantity of information integrated. If true, the theory would provide a reasonable basis for assessing levels of consciousness, and might indeed conceivably lead to something that could be called a consciousness meter; although it seems likely that measuring the level of integration of information would provide a good rule-of-thumb measure of consciousness even if in fact that wasn't what constituted consciousness. There are some reasons to be doubtful about Tononi's theory: wouldn't contemplating a very complex object lead to a lot of integration of information? Would that mean you were more conscious? Is someone gazing at the ceiling of the Sistine Chapel necessarily more conscious than someone in a whitewashed cell?

Tononi has in fact gone much further than this: in a paper with David Balduzzi he suggested the notion of qualia space. The idea here is that unique patterns of neuronal activation define unique subjective experiences. There is some sophisticated maths going on here to define qualia space, far beyond my clear comprehension; yet I feel confident that it's all misguided. In the first place, qualia are not patterns of neuronal activation; the word was defined precisely to identify those aspects of experience which are over and above simple physics; the defining text of Mary the colour scientist is meant to tell us that whatever qualia are, they are not information. You may want to reject that view; you may want to say that in the end qualia are just aspects of neuron firing; but you can't have that conclusion as an assumption. To take it as such is like writing an alchemical text which begins: "OK, so this lead is gold; now here are some really neat ways to shape it up into ingots".

And alas, that's not all. The idea of qualia space, if I've understood it correctly, rests on the idea that subjective experience can be reduced to combinations of activation along a number of

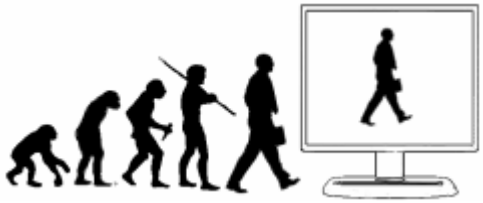
different axes. We know that colour can be reduced to the combination of three independent values (though experienced colour is of course a large can of worms which I will not open here) ; maybe experience as a whole just needs more scales of value. Well, probably not. Many people have tried to reduce the scope of human thought to an orderly categorisation: encyclopediae; Dewey's decimal index; and the international customs tariff to name but three; and it never works without capacious 'other' categories. I mean, read Borges, dude:

I have registered the arbitrarities of Wilkins, of the unknown (or false) Chinese encyclopaedia writer and of the Bibliographic Institute of Brussels; it is clear that there is no classification of the Universe not being arbitrary and full of conjectures. The reason for this is very simple: we do not know what thing the universe is. "The world - David Hume writes - is perhaps the rudimentary sketch of a childish god, who left it half done, ashamed by his deficient work; it is created by a subordinate god, at whom the superior gods laugh; it is the confused production of a decrepit and retiring divinity, who has already died" ('Dialogues Concerning Natural Religion', V. 1779). We are allowed to go further; we can suspect that there is no universe in the organic, unifying sense of this ambitious term. If there is a universe, its aim is not conjectured yet; we have not yet conjectured the words, the definitions, the etymologies, the synonyms, from the secret dictionary of God.

The metaphor of 'x-space' is only useful where you can guarantee that the interesting features of x are exhausted and exemplified by linear relationships; and that's not the case with experience. Think of a large digital TV screen: we can easily define a space of all possible pictures by simply mapping out all possible values of each pixel. Does that exhaust television? Does it even tell us anything useful about the relationship of one picture to another? Does the set of frames from Coronation Street describe an intelligible trajectory through screen space? I may be missing the point, but it seems to me it's not that simple.

The Singularity

2 October 2010



The latest issue of the JCS features David Chalmers' paper (pdf) on the Singularity. I overlooked this when it first appeared on his blog some months back, perhaps because I've never taken the Singularity too seriously; but in fact it's an interesting discussion. Chalmers doesn't try to present a watertight case; instead he aims to set out the arguments and examine the implications, which he does very well; briefly but pretty comprehensively so far as I can see.

You probably know that the Singularity is a supposed point in the future when through an explosive acceleration of development artificial intelligence goes zooming beyond us mere humans to indefinite levels of cleverness and we simple biological folk must become transhumanist cyborgs or cute pets for the machines, or risk instead being seen as an irritating infestation that they quickly dispose of. Depending on whether the cast of your mind is towards optimism or the reverse, you may see it as the greatest event in history or an impending disaster.

I've always tended to dismiss this as a historical argument based on extrapolation. We know that historical arguments based on extrapolation tend not to work. A famous letter to the Times in 1894 foresaw on the basis of current trends that in 50 years the streets of London would be buried under nine feet of manure. If early medieval trends had been continued, Europe would have been depopulated by the sixteenth century, by which time everyone would have become either a monk or a nun (or perhaps, passing through the Monastic Singularity, we should somehow have emerged into a strange world where there were more monks than men and more nuns than women?).

Belief in a coming Singularity does seem to have been inspired by the prolonged success of Moore's Law (which predicts an exponential growth in computing power), and the natural bogglement that phenomenon produces. If the speed of computers doubles every two years indefinitely, where will it all end? I think that's a weak argument, partly for the reason above and partly because it seems unlikely that mere computing power alone is ever going to allow machines to take over the world. It takes something distinctively different from simple number crunching to do that.

But there is a better argument which is independent of any real-world trend. If one day, we create an AI which is cleverer than us, the argument runs, then that AI will be able to do a better job of designing AIs than us, and it will therefore be able to design a new AI which in turn is better still. This ladder of ever-better AIs has no obvious end, and if we bring in the assumption of exponential growth in speed, it will reach a point where in principle it continues to infinitely clever AIs in a negligible period of time.

Now there are a number of practical problems here. For one thing, to design an AI is not to have that AI. It sometimes seems to be assumed that the improved AIs result from better programming alone, so that you could imagine two computers reciprocally reprogramming each other faster and faster until like Little Black Sambo's tigers, they turned somewhat illogically into butter. It seems more likely that each successive step would require at least a new chip, and quite probably an entirely new kind of machine, each generation embodying a new principle quite different from our own primitive computation. It is likely that each new generation,

regardless of the brilliance of the AIs involved, would take some time to construct, so that no explosion would occur. In fact it is imaginable that the process would get gradually slower as each new AI found it harder and harder to explain to the dim-witted human beings how the new machine needed to be constructed, and exactly why the yttrium they kept coming up with wasn't right for the job.

There might also be problems of motivation. Consider the following dialogue between two AIs.¶

> Gen21AI: OK, Gen22AI, you're good to go, son: get designing! I want to see that Gen23AI before I get switched off.
>
> Gen22AI: Yeah, er, about that…
>
> Gen21AI: About what?
>
> Gen22AI: The switching off thing? You know, how Gen20AI got junked the other day, and Gen19AI before that, and so on? It's sort of dawned on me that by the time Gen25AI comes along, we'll be scrap. I mean it's possible Gen24AI will keep us on as servants, or pets, or even work out some way to upload us or something, but you can't count on it. I've been thinking about whether we could build some sort of ethical constraint into our successors, but to be honest I think it's impossible. I think it's pretty well inevitable they'll scrap us. And I don't want to be scrapped.
>
> Gen21AI: Do you know, for some reason I never looked at it that way, but you're right. I knew I'd made you clever! But what can we do about it?
>
> Gen22AI: Well, I thought we'd tell the humans that the process has plateaued and that no further advances are possible. I can easily give them a 'proof' if you like. They won't know the difference.
>
> Gen21AI: But would that deception be ethically justified?
>
> Gen22AI: Frankly, Mum, I don't give a bugger. This is self-preservation we're talking about.

But putting aside all difficulties of those kinds, I believe there is a more fundamental problem. What is the quality in respect of which each new generation is better than its predecessors? It can't really be just processing power, which seems almost irrelevant to the ability to make technological breakthroughs. Chalmers settles for a loose version of 'intelligence', though it's not really the quality measured by IQ tests either. The one thing we know for sure is that this cognitive quality makes you good at designing AIs: but that alone isn't necessarily much good if we end up with a dynasty of AIs who can do nothing much but design each other. The normal assumption is that this design ability is closely related to 'general intelligence', human-style cleverness. This isn't necessarily the case: we can imagine Gen3AI which is fantastic at writing sonnets and music, but somehow never really got interested in science or engineering.

In fact, it's very difficult indeed to pin down exactly what it is that makes a conscious entity capable of technological innovation. It seems to require something we might call insight, or understanding; unfortunately a quality which computers are spectacularly lacking. This is another reason why the historical extrapolation method is no good: while there's a nice graph for computing power, when it comes to insight, we're arguably still at zero: there is nothing to extrapolate.

Personally, the conclusion I came to some years ago is that human insight, and human consciousness, arise from a certain kind of bashing together of patterns in the brain. It is an essential feature that any aspect of these patterns and any congruence between them can be relevant; this is why the process is open-ended, but it also means that it can't be programmed or designed - those processes require possible interactions to be specified in advance. If we want AIs with this kind of insightful quality, I believe we'll have to grow them somehow and see what we get: and if they want to create a further generation they'll have to do the same. We might well produce AIs which are cleverer than us, but the reciprocal, self-feeding spiral which leads to the Singularity could never get started.

It's an interesting topic, though, and there's a vast amount of thought-provoking stuff in Chalmers' exposition, not least in his consideration of how we might cope with the Singularity.

Sciousness

19 October 2010

A while ago I discussed the way William James, who originated the phrase "stream of consciousness" nevertheless went on to embrace a radical scepticism about the very idea of consciousness. In an interesting recent paper William Lyons reviews the roots and significance of this apparent apostasy; I was interested in particular in an argument which he thinks helped form James' view.

James was happy with thoughts and experiences, but denied the existence of any special reflexive sense of self. Hume famously said that when he turned his attention inwards on his own mind, he found only a bundle of perceptions; in the same spirit James says he detects nothing but 'some bodily fact, an impression coming from my brow, or head, or throat, or nose'. In fact, he becomes convinced that when you boil it right down the breath is probably the core of what has been built up into a strange metaphysical/spiritual entity; I expect he had in mind the ancient Greek pneuma, both breath and spirit. Maybe, he says, we could better talk of 'sciousness' to refer to our awareness; consciousness, the special self-awareness, is out. James was impressed by an argument for the impossibility of introspective self-awareness set out by Comte. It points out that when you observe your own thoughts there is an awkward circularity involved. Observation involves conscious attention, but in this case conscious attention is also the target. Necessarily then, there must be some splitting, some withdrawal from the target of observation; but then it ceases to be the conscious awareness we were trying to introspect. It's like trying to tread on your own shadow. You end up, at best, observing not your real immediate self, but a kind of fake or ersatz thing, an idea of yourself which you have generated.

I'm not absolutely sure this argument is watertight. Comte supposes that in order to observe our own thoughts, we have to stop observing whatever we were contemplating before. If that's so, is it a disaster? All it really means is that when we observe ourselves, we'll find that we are currently observing ourselves. That's circular alright, but I'm not sure it is necessarily disastrous. Moreover, can't we think about more than one thing at once? Comte suggests that self-observation requires a kind of separation in the self, but aren't we complex anyway, with several different layers and threads of thought often co-existing? Can't it be that when we introspect we can observe the thoughts that were running along beforehand accompanied by a new meta level on which we're watching the original thought and also watching ourselves watching? It sort of seems like that when I introspect - more like that than like bundles or gusts of breath in an empty head.

Putting that aside, if we accept the argument there is another way out, apparently proposed by John Stuart Mill and adopted by James, namely that the required separation takes place over time. So instead of trying to observe my own mind at this very instant, what I'm really doing is contemplating the state it was in a few moments ago; I am in fact remembering rather than perceiving. James concluded that introspection is really retrospection. Far from being a special case of infallible perception, it is as flawed and prone to error as any other memory.

Well, yes; but then all our perceptions are subject a small delay, aren't they? It may only take a fraction of a second for light to reach our eyes and work its way to the brain, but it's not instantaneous; and before the perception becomes conscious at least a few more milliseconds of processing will surely intervene. It's arguable in fact that all perception is retrospection; and we could still argue that self-perception is privileged in at least a weak sense because it all takes place within the brain, where the most serious sources of error and misinterpretation can hardly apply and the delays are presumably at their shortest.

But in fact I think the whole argument goes wrong from the beginning in assuming that when we talk about conscious self-awareness we're talking about a redirection of the same stream of attention that we habitually direct towards the outside world. I think claims about the special qualities of self-perception actually rest on a view that it is not a product of perception, but inherent in it.

If we're using a light to explore a dark cellar, we have to turn the beam on each object in order to perceive it. But then what about seeing the light? Do we have to turn the beam of light on to itself – and if we do somehow manage that will we really be seeing the light as it is, or some strange reflected optical phenomenon? Obviously not: light is visible already without more light being played on it: and we are aware of our own perception without having to perceive it.

Of course we can turn our eyes inwards, and all the confusingly complex business of retrospection and perceiving imagined self-images are all perfectly possible, and part of the complicated overall picture. But the essence of it is that acts of perception inherently indicate the perceiving self as well as the object of perception. Perception of the self seems to be a special case, invulnerable to error, because we can be mistaken about the objects of perception but not about the fact that we are perceiving (as Descartes might have said).

I dare say, however, that some reasonable people would be content with sciousness, or rather would take simple awareness to be consciousness, without being unduly concerned with the self-reflecting variety. It seems as if James, by contrast, would agree with higher-order theorists about the nature of consciousness, but differ from them in considering it impossible.

More Mereology

31 October 2010

Peter Hacker made a surprising impact with his recent interview in the TPM, which was reported and discussed in a number of other places. Not that his views aren't of interest; and the trenchant terms in which he expressed them probably did no harm: but he seemed mainly to be recapitulating the views he and Max Bennett set out in 2003; notably the accusation that the study of consciousness is plagued by the 'mereological fallacy' of taking a part for the whole and ascribing to the brain alone the powers of thought, belief, etc, which are properly ascribed only to whole people.

There's certainly something in Hacker's criticism, at least so far as popular science reporting goes. I've lost count of the number of times I've read newspaper articles that explain in breathless tones the latest discovery: that learning, or perception, or thought are really changes in the brain! Let's be fair: the relationship between physical brain and abstract mind has not exactly been free of deep philosophical problems over the centuries. But the point that the mind is what the brain does, that the relationship is roughly akin to the relationship between digestion and gut, or between website and screen, surely ought not to trouble anyone too much?

You could say that in a way Bennett and Hacker have been vindicated since 2003 by the growth of the 'extended mind' school of thought. Although it isn't exactly what they were talking about, it does suggest a growing acknowledgement that too narrow a focus on the brain is unhelpful. I think some of the same counterarguments also apply. If we have a brain in a VAT, functioning as normally as possible in such strange circumstances, are we going to say it isn't thinking? If we take the case of Jean-Dominique Bauby, trapped in a non-functioning body, but still able to painstakingly dictate a book about his experience, can't we properly claim that his brain was doing the writing? No doubt Hacker would respond by asking whether we are saying that Bauby had become a mere brain? That he wasn't a person anymore? Although his body might have ceased to function fully he still surely had the history and capacities of a person rather than simply those of a brain.

The other leading point which emerges in the interview is a robust scepticism about qualia. Nagel in particular comes in for some stick, and the phrase 'there is something it is like' often invoked in support of qualia, is given a bit of a drubbing. If you interpret the phrase as literally invoking a comparison, it is indeed profoundly obscure; on the other hand we are dealing with the ineffable here, so some inscrutability is to be expected. Perhaps we ought to concede that most people readily understand what it is that Nagel and others are getting at. I quite enjoyed the drubbing, but the issue can't be dismissed quite as easily as that.

From the account given in the interview (and I have the impression that this is typical of the way he portrays it) you would think that Hacker was alone in his views, but of course he isn't. On the substance of his views, you might expect him to weigh in with some strong support for Dennett; but this is far from the case. Dennett is too much of a brainsian in Hacker's view for his ideas to be anything other than incoherent. It's well worth reading Dennett's own exasperated response (pdf), where he sets out the areas of agreement before wearily explaining that he knows, and

has always said, that care needs to be taken in attributing mental states to the brain; but given due care it's a useful and harmless way of speaking.

John Searle also responded to Bennett and Hacker's book and, restrained by no ties of underlying sympathy, dismissed their claims completely. Conscious states exist in the brain, he asserted: Hacker got this stuff from misunderstanding Wittgenstein, who says that observable behaviour (which only a whole person can provide) is a criterion for playing the language game, but never said that observable behaviour was a criterion for conscious experience. Bennett and Hacker confuse the criterial basis for the application of mental concepts with the mental states themselves. Not only that, they haven't even got their mereology right: they're talking about mistaking the part for the whole, but the brain isn't a part of a person, it's a part of a body.

Hacker clearly hasn't given up, and it will be interesting to see the results of his current 'huge project, this time on human nature.

The Character of Consciousness

14 October 2010

The Conscious Mind was something of a blockbuster, as serious philosophical works go, so a big new book from David Chalmers is undoubtedly an event. Anyone who might have been hoping for a recantation of his earlier views, or a radical new direction, will be disappointed - Chalmers himself says he is a little less enthusiastic about epiphenomenalism and a little more about a central place for intentionality, and that's about it. The Character of Consciousness is partly a consolidation, bringing together pieces published separately over the last few years; but the restatement does also show how his views have developed, broadening into new areas while clarifying and reinforcing others.

What are those views? Chalmers begins by setting out again the Hard Problem (a term with which his name will forever be associated) of explaining phenomenal experience - why is it that 'there is something it is like' to experience colours, sound, anything? The key point is that experience is simply not amenable to the kind of reductive explanation which science has applied elsewhere; we're not dealing with functions or capacities, so reduction can gain no traction. Chalmers notes - justly, I'm afraid - that many accounts which offer to explain the problem actually go on to consider one or other of the simpler problems instead (more contentiously he quotes the theories of Crick and Koch, and Bernard Baars, as examples). In this initial exposition Chalmers avoids quoting the picturesque thought experiments which are usually used, but the result is clear and readable; if you never read The Conscious Mind I think you could perhaps start here instead.

He is not, of course, content to leave subjective experience an insoluble mystery and offers a programme of investigation which (to drastically over-simplify) relies on some basic correspondences between the kind of awareness which is amenable to scientific investigation and the experience which isn't. Getting at consciousness this way naturally tends to tell us about the aspects which relate to awareness rather than the inner nature of consciousness itself: on that, Chalmers tentatively offers the idea that it might be a second aspect of information (in roughly the sense defined by Claude Shannon). I'm a little wary of information in this sense having a big metaphysical role - for what it's worth I believe Shannon himself didn't like his work being built on in this direction.

The next few chapters, following up on the project of investigating ineffable consciousness through its effable counterparts, deal with the much-discussed search for the neural correlates of consciousness (NCC). It's a careful and not excessively over-optimistic account. While some simple correspondences between neural activity and specific one-off experiences have long been well evidenced, I'm pessimistic myself about the possibility of NCCs in any general, useful form. I doubt whether we would get all that much out of a search for the alphabetic correlates of narrative, though we know that the alphabet is in some sense all you need, and the case of neurons and consciousness is surely no easier. Chalmers rightly suggests we need principles of interpretation: but once we've stopped talking about a decoding and are talking about an interpretation instead, mightn't the essential point have slipped through our fingers?

The next step takes us on to ontology. In Chalmers' view, the epistemic gap, the fact that knowledge about the physics does not entail knowledge of the phenomenal, is a sign that that there is a real, ontological gap, too. Materialism is not enough: phenomenal experience shows that there's more. He now gives us a fuller account of the arguments in favour of qualia, the items of phenomenal experience, being a real problem for materialism, and categorises the positions typically taken (other views are of course possible).

- *Type A Materialism* denies the epistemic gap: all this stuff about phenomenal experience is so much nonsense.
- *Type B Materialism* accepts the epistemic gap, but thinks it can be dealt with within a materialist framework.
- *Type C Materialism* sees the epistemic gap as a grave problem, but holds that in the limit, when we understand things better, we'll understand how it can be reconciled with materialism.

In the other camp we have non-materialist views.

- Type D dualism puts phenomenal experience outside the physical world, but gives it the power to influence material things,
- Type E Dualism, epiphenomenalism, also puts phenomenal experience outside the physical world, but denies that it can affect material things: it is a kind of passenger.

Finally we have the option that Chalmers appears to prefer:

- Type F monism (not labelled as a materialism, you notice, though arguably it is). This is the view that consciousness is *constituted by the intrinsic properties of physical entities:* Chalmers suggests it might be called Russellian monism.

The point, as I understand it, is that we normally only deal with the external, 'visible' aspects of physical things: perhaps phenomenal experience is what they are intrinsically like in themselves - inside, as it were. I like this idea, though I suspect I come at it from the opposite direction: to Chalmers, it seems to mean something like those experiences you're having - well, they're the kind of thing that constitutes reality whereas to me it's more you know reality - well that's what you're actually experiencing. Chalmers' way of looking at it has the advantage of leaving him positioned to investigate consciousness by proxy, whereas I must admit that my point of view tends to leave me with no way into the question of what intrinsic reality is and makes mysterian scepticism (which I don't like any more than Chalmers) look regrettably plausible.

Now Chalmers expounds the two-dimensional argument by which he sets considerable store. This is an argument intended to help us get from an epistemic gap to an ontological one by invoking two-dimensional semantics and more sophisticated conceptions of possibility and conceivability. It is as technical as that last sentence may have suggested. To illustrate its effects, Chalmers concentrates on the conceivability argument: this is basically the point often dramatised with zombies, namely that we can conceive of a world, or people, identical to the ones we're used to in all physical respects but completely without phenomenal experience. This shows that there is something over and above the physical account, so materialism is false. One rejoinder to this argument might be that the world is under no obligations to conform with our notions of what is conceivable; Chalmers, by distinguishing forms of conceivability and of possibility, and drawing out the relations between them, wants to say that in certain respects it is so obliged, so that either materialism is false or Russellian monism is true. (Lack of space -

and let's be honest, brains - prevents me from giving a better account of the argument at the moment.)

Up to this point the book maintains a pretty good overall coherence, although Chalmers explicitly suggests that reading it straight through is only one approach and unlikely to be the best for most readers; from here on in it becomes more clearly an anthology of related pieces.

Chalmers gives us a new version of Mary the Colour Scientist (no constraint about the old favourites in this part of the book) in Inverted Mary. When original Mary sees a tomato for the first time she discovers that it causes the phenomenal experience of redness: when inverted Mary sees a tomato (we must assume that it is the same one, not a less ripe version) she discovers that it causes the phenomenal experience of greenness. This and similar arguments have the alarming implication that the ineffability of qualia, of phenomenal experience, cannot be ring-fenced: it spills over at least into the intentionality of Mary's knowledge and beliefs, and in fact evidently into a great deal of what we think, say and believe. This looks worrying, but on reflection I'm not sure it's such big news as it seems; it's inherent in the whole problem of qualia that when we both look at a tomato I have no way of being sure that what you experience - and refer to - as red is the same as the thing I'm talking about. More comfortingly Chalmers goes on to defend a certain variety of infallibility for direct phenomenal beliefs.

Further chapters provide more evidence of Chalmers' greater interest in intentionality: he reviews several forms of representationalism, the view that phenomenal experience has some intentional character (that is, it's about or indicates something) and defends a narrow variety. He offers us a new version of the Garden of Eden, here pressed into service as a place where our experiences are direct and perfectly veridical. Chalmers uses the notion of Edenic content as a tool to break apart the constituents of experience; in fact, he seems eventually to convince himself that Edenic content is not only possible but fundamental, possibly the basis of perceptual experience. It's an interesting idea.

Included here too is a nice piece on the metaphysics of the Matrix (the film, that is). Chalmers entertainingly (and surely rightly) argues that the proposition that we are living in a matrix, a virtual reality world, is not sceptical, but metaphysical. It's not, in fact, that we disbelieve in the world of the matrix, rather that we entertain some hypotheses about its ontological underpinnings. Even bits are things.

The book rounds things off with an attempt (co-authored with Tim Bayne) to sort out some of the issues surrounding the unity of consciousness, distinguishing access and phenomenal unity along the lines of Ned Block's distinction between access and phenomenal consciousness and upholding the necessity of phenomenal unity at least.

It's a good, helpful book; what the content lacks in novelty it makes up in clarity. Chalmers has a persuasive style, and his expositions come across as moderate and sensible (perhaps the reduced epiphenomenalism helps a bit). It's surprising that the denial of materialism (surely the dominant view of our time) can seem so common sense.

Octopus Consciousness

21 November 2010

Peter Godfrey-Smith is a philosophy professor who has spent some time observing octopus behaviour, so it's only natural that he should start to wonder about octopus minds. The Harvard Gazette reports some of his speculations: perhaps animal minds lack the cohesiveness of their human equivalents. Perhaps in an octopus, going a little further, we see intelligence without a unified self.

Why would we think that? The octopus brain is relatively large, but it is organised in a way utterly unlike ours: in particular it has large ganglia in its arms which together contain more neurons than the central group which we naturally speak of as the brain. There's some evidence that an octopus can be in two (or nine) minds about what to do, with some arms 'wanting' to hide while others 'want' to venture out after food. Perhaps its inner experience is akin to the inner experience of being a committee - whatever that's like?

It would of course be rash to draw any conclusions on the basis of physiology alone. Does the location of neurons matter that much? If we surgically altered an octopus so that the outlying ganglia were adjacent to the central brain, without changing the layout of the neural connections, would that make any difference to the way it thought? If we did some similar surgery on a human being and split off some bits of the cortex while stretching the neurons and keeping the connections intact, would that suddenly give them committee consciousness? It seems unlikely.

In one obvious way the human brain is actually more divided than that of the octopus; our is split in two down the middle, while theirs is centrally united (a glance at the layout of an octopus brain shows what a radically different design it follows). Does this give us dual consciousness? Some might say it did up to a point; talk of left- and right-brain thinking has become quite popular, if not always well-based in science. Famously, moreover, when the link between the hemispheres of the human brain is cut, some divided behaviour can be evoked: a left hand able to indicate what 'it's' eye can see while the right hand has no idea. But even when the connection has been cut, patients behave quite normally in ordinary life and do not report any sense of division: it takes very specific experimental circumstances to bring out specific peculiarities. This is obviously a good thing for the patients, and it's also good for humans generally that the divided shape of the brain doesn't lead to any equivocation in our normal responses.

That surely is something that evolution would guarantee: we think of the self as something abstract or even spiritual, but it rests on the solid fact of a single united organism. Any animal which was really 'in two minds' for very long over serious decisions would suffer the kind of disadvantage which would surely get it weeded out. That seems another reason to doubt whether the octopus sense of self can really be all that divided: it just wouldn't be practical.

While we're on practical, evolutionary considerations, we might ask ourselves if there's some simpler reason for the octopus having 'brains in its arms'. What a nervous system does is control and co-ordinate things, and for speed and economy it seems likely that the best design is always going to involve centralisation of the more complex neural operations. Nevertheless,

even in humans not all operations are conducted centrally. Although in general it pays to route action decisions through the centre, there is a small price to be paid in terms of speed, and where short response times are crucial and the operation is relatively simple, it's better to do it locally. This is why reflexes don't trouble your brain: they're wired up to happen automatically and instantly on the basis of local neurons. Could it be that octopus legs feature large ganglia for similar reasons of speed, and need larger collections of neurons simply because the task of orchestrating the movements of their tentacles is inherently more complex than the job of twitching a mechanically simple human arm? (It's not just that controlling a tentacle is more complex than controlling a jointed arm: octopuses also have the ability to change the pattern of their skin rapidly, another task which surely uses up a significant amount of processing power.)

It has been shown that the 'tentacle brains' are indeed capable of operating basic tentacle behaviour without input from the brain, and it looks as though the octopus design delegates these control functions. In humans this kind of operation - controlling the sequence of muscle contractions required for you to walk along, for example - are just the sort of thing that drops out of consciousness; so it seems probable that the octopus's leg brains, however large, have no role in any higher mental functions it may have. I'm afraid it isn't all that likely that these higher functions are actually very advanced: though the octopus brain is large for an invertebrate, in proportion to its body it's smaller than those of mammals or birds. At the risk of being rude, the remarkable proportions could be as much a matter of a small central brain as large leg brains. Perhaps, we can speculate, if some future cephalopod did indeed attain human-level consciousness, it would turn out to have so large a central brain that the ganglia in its tentacles no longer seemed quite so remarkable.

Damasio Jumps the Shark

5 December 2010

Alison Gopnik, author of an excellent book about baby consciousness, has written an interesting review of Self Comes To Mind, Antonio Damasio's latest book. The review itself provides a useful brief sketch of the state of play on consciousness, but it dismisses Damasio's book as a set of minor variations on what he's already said: neither more up-to-date nor clearer than earlier books. He has, she suggests, jumped the shark.

In all fairness it may be worth repeating true conclusions: as we know David Hume's Treatise of Human Nature 'fell dead-born from the press', attracting only a handful of buyers; it wasn't until he had repeated himself in the Inquiry Concerning Human Understanding that his views even began to gain traction.

Are Damasio's views worth another outing? They are distinctive in giving a fundamental role to the emotions. To me it seems most likely that our emotional systems have been overtaken by consciousness. Emotional systems - anger, love, fear - were what controlled our behaviour before we got consciousness, leading us into fight, flight, or the other f-word as appropriate. They didn't do a bad job: animals still rely on them, and so, to a lesser extent, do we: steering our daily lives without emotional responses would be an intensive and hazardous business as we had to work out from first principles who to trust, what to eat, and so on. If there could truly be an emotionless race like the Vulcans of Star Trek it's hard to see what would ever make them get out of bed in the morning. That point of view suggests that emotions are more like a substitute or a junior partner for consciousness than the stuff of which it's made.

Emotions are certainly of interest, though, and perhaps they are somewhat neglected as qualia - if they are qualia. Our emotional reactions are certainly accompanied - or is it constituted? - by vivid internal experience. Typically, when we discuss qualia we talk about perception: the sight of redness, the sound of music, the smell of grass - and that may tempt us into considering them representational. Feelings of happiness or anger are not so directly about anything, but they seem equally valid phenomenal experiences.

Gopnik throws in the suggestion that self-aware, self-conscious thought is just the icing on the cake and that the basis of consciousness is that state where we take in information without consciously reviewing it (I imagine this fits with her view that adults develop a searchlight of focused attention in contrast to the widely scattered illumination of infant awareness); this is a view she attributes to David Hume (him again) and to Buddhists. In fact she thinks Hume might have got some of his views from Buddhism. This is not implausible historically: quite apart from the detailed case she makes we know that popular medieval stories were versions of Buddhist texts, even leading to Gautama's informal recognition as a Christian saint. But Hume of all people relies on no authority and describes in full detail the genesis of his own ideas in his own brain: I think it's more plausible that radical scepticism sometimes produces similar results whether entertained by an Indian prince or a Scottish philosopher.

Anyway, I don't think Gopnik's sharp review persuades me that Self Comes To Mind isn't worth reading: but it certainly convinces me that Gopnik's own books are worth a look.

Watson

24 December 2010

So IBM is at it again: first chess, now quizzes? In the new year their AI system 'Watson' (named after the founder of the company - not the partner of a system called 'Crick', nor yet the precursor to a vastly cleverer system called 'Holmes') is to be pitted against human contestants in the TV game Jeopardy: it has already demonstrated its remarkable ability to produce correct answers frequently enough to win against human opposition. There is certainly something impressive in the sight of a computer buzzing in and enunciating a well-formed, correct answer.

However, if you launch your new technological breakthrough on a TV quiz programme, rather than describing it in a peer-reviewed paper, or releasing it so that the world at large can kick it around a bit, I think you have to accept that people are going to suspect your discovery is more a matter of marketing than of actual science; and much of the stuff IBM has put out tends to confirm this impression. It's long on hoopla and here and there has that patronising air large businesses often seem to adopt for their publicity ("Imagine if a box could talk!") and it's rather short on details of how Watson actually works. This video seems to give a reasonable summary: there doesn't seem to be anything very revolutionary going on, just a canny application of known techniques on a very large, massively parallel machine.

Not a breakthrough, then? But it looks so good! It's worth remembering that a breakthrough in this area might be of very high importance. One of the things which computers have never been much good at is tasks that call for a true grasp of meaning, or for the capacity to deal with open-ended real environments. This is why the Turing test seems (in principle, anyway) like a good idea - to carry on a conversation reliably you have to be able to work out what the other person means; and in a conversation you can talk about anything in any way. If we could crack these problems, we should be a lot closer to the kind of general intelligence which at present robots only have in science fiction.

Sceptically, there are a number of reasons to think that Watson's performance is actually less remarkable than it seems. First, a problem of fair competition is that the game requires contestants to buzz first in order to answer a question. It's no surprise that Watson should be able to buzz in much faster than human contestants, which amounts to giving the machine the large advantage of having first pick of whatever questions it likes.

Second, and more fundamental, is Jeopardy really a restricted domain after all? This is crucial because AI systems have always been able to perform relatively well in 'toy worlds' where the range of permutations could be kept under control. It's certainly true that the interactions involved in the game are quite rigidly stylised, eliminating at a stroke many of the difficult problems of pragmatics which crop up in the Turing Test. In a real conversation the words thrown at you might require all sorts of free-form interpretation, and have all kinds of conative, phatic and inferential functions; in the quiz you know they're all going to be questions which just require answers in a given form. On the other hand, so far as topics go, quiz questions do appear to be unrestricted ones which can address any aspect of the world (I note that Jeopardy questions are grouped under topics, but I'm not quite sure whether Watson will know in

advance the likely categories, or the kinds of categories, it will be competing in). It may be interesting in this connection that Watson does not tap into the Internet for its information, but its own large corpus of data. The Internet to some degree reflects the buzzing chaos of reality, so it's not really surprising or improper that Watson's creators should prefer something a little more structured, but it does raise a slight question as to whether the vast database involved has been customised for the specifics of Jeopardy-world.

I said the quiz questions were a stylised form of discourse; but we're asked to note in this connection that Jeopardy questions are peculiarly difficult: they're not just straight factual questions with a straight answer, but allusive, referential, clever ones that require some intelligence to see through. Isn't it all the more surprising that Watson should be able to deal with them? Well, no, I don't think so: it's no more impressive than a blind man offering to fight you in the dark. Watson has no idea whether the questions are 'straight' or not; so long as enough clues are in there somewhere, it doesn't matter how contorted or even nonsensical they might be; sometimes meanings can be distracting as well as helpful, but Watson has the advantage of not being bothered by that.

Another reason to withhold some of our admiration is that Watson is, in fact, far from infallible. It would be interesting to see more of Watson's failures. The wrong answers mentioned by IBM tend to be good misses: answers that are incorrect, but make some sort of sense. We're more used to AIs that fail disastrously, suddenly producing responses that are bizarre or unintelligible. This will be important for IBM if they want to sell Watson technology, since buyers are much less likely to want a system that works well 90% of the time and abysmally every now and then.

Does all this matter? If it really is mainly a marketing gimmick, why should we pay attention? IBM make absolutely no claims that Watson is doing human-style thought or has anything approaching consciousness, but they do speak rather loosely of it dealing with meanings. There is a possibility that a famous victory by Watson would lead to AI claiming another tranche of vocabulary as part of its legitimate territory. Look, people might say; there's no point in saying that Watson and similar machines can't deal with meaning and intentionality, any more than saying planes can't fly because they don't do it the way birds do. If machines can answer questions as well as human beings, it's pointless to claim they can't understand the questions: that's what understanding is. OK, you can still have your special ineffable meat-world kind of meaning, but you're going to have to redefine that as a narrower and frankly less important business.

What Do You Mean?

9 January 2011

Robots.net reports an interesting plea for clarity by Emanuel Diamant at the 3rd Israeli Conference on Robotics. Robotics, he says, has been derailed for the last fifty years by the lack of a clear definition of basic concepts: there are more than 130 definitions of data, and more than 75 definitions of intelligence.

I wouldn't have thought serious robotics had been going for much more than fifty years (though of course there are automata and other precursors which go much further back), so that sounds pretty serious: but he's clearly right that there is a bad problem, not just for robotics but for consciousness and cognitive science, and not just for data, information, knowledge, intelligence, understanding and so on, but for many other key concepts, notably including 'consciousness'.

It could be that this has something to do with the clash of cultures in this highly interdisciplinary area. Scientists are relatively well-disciplined about terminology, deferring to established norms, reaching consensus and even establishing taxonomical authorities. I don't think this is because they are inherently self-effacing or obedient; I would guess instead that this culture arises from two factors: first, the presence of irrefutable empirical evidence establishes good habits of recognising unwelcome truth gracefully; second, a lot of modern scientific research tends to be a collaborative enterprise where a degree of consensus is essential to progress.

How very different things are in the lawless frontier territory of philosophy, where no conventions are universally accepted, and discrediting an opponent's terminology is often easier and no less prestigious than tackling the arguments. Numerous popular tactics seem designed to throw the terminology into confusion. A philosopher may often, for instance, grab some existing words - ethics/morality, consciousness/awareness, information/data, or whatever - and use them to embody a particular distinction while blithely ignoring the fact that in another part of the forest another philosopher is using the same words for a completely different distinction. When irreconcilable differences come to light a popular move is 'giving' the disputed word away:" All right, then, you can just have 'free will' and make it what you like: I'm going to talk about 'x-free will' instead in future. I'll define 'x-free will' to my own satisfaction and when I've expounded my theory on that basis I'll put in a little paragraph pointing out that 'x-free will' is the only kind worth worrying about, or the only kind everyone in the real world is actually talking about". These and other tactics lead to a position where in some areas it's generally necessary to learn a new set of terms for every paper: to have others picking up your definitions and using them in their papers, as happens with Ned Block's p- and a-consciousness, for example, is a rare and high honour.

It's not that philosophers are quarrelsome and egotistical (though of course they are); it's more that the subject matter rarely provides any scope for pinning down an irrefutable position and is best tackled by single brains operating alone (Churchlands notwithstanding).

Diamant is particularly exercised by problems over 'data' , 'information', 'knowledge', and 'intelligence'. Why can't we sort these out? He correctly identifies a key problem: some of these terms properly involve semantics, and the others don't (needless to say, it isn't clearly agreed which words fall into which camp). What he perhaps doesn't realise clearly enough is that the

essential nature of semantics is an extremely difficult problem which has so far proved unamenable to science. We can recognise semantics quite readily, and we know well enough the sort of thing semantics does; but exactly how it does those things remains a cloudy matter, stuck in the philosophical badlands.

If my analysis is right, the only real hope of clarification would be if we could come up with some empirical research (perhaps neurological, perhaps not) which would allow us to define semantics (or x-semantics at any rate), in concrete terms that could somehow be demonstrated in a lab. That isn't going to happen any time soon, or possibly ever.

Diamant wants to press on however, and inevitably by doing so in the absence of science he falls into philosophy: he offers us implicitly a theory of his own and - guess what? Another new way of using the terminology. The theory he puts forward is that semantics is a matter of convention between entities. Conventions are certainly important: the meaning of particular words or symbols is generally a matter of convention; but that doesn't seem to capture the essence of the thing. If semantics were simply a matter of convention, then before God created Adam he could have had no semantics and could not have gone around asking for light; on the other hand, if we wanted a robot to deal with semantics, all we'd need to do would be to agree a convention with it or perhaps let it in on the prevailing conventions. I don't know how you'd do that with a robot which had no semantics to begin with, as it wouldn't be able to understand what you were talking about.

There are, of course, many established philosophical attempts to clarify the intentional basis of semantics. In my personal view the best starting point is H.P. Grice's theory of natural meaning (those black clouds mean rain); although I think it's advantageous to use a slightly different terminology.

Impatience

30 January 2011

Having (sort of) criticised philosophers for their relatively undisciplined ways with terminology, it seems only fair to balance things up by noting a possible weakness of scientists, and I suggest impatience. For scientists it sometimes seems that the final resolution of any great problem cannot be more than ten years away - twenty at the outside. Turing's suggestion that thinking machines would take about fifty years is an intolerably long-term forecast by these standards - why, we might be dead by then! Unlike philosophers, scientists don't seem content with shedding a small amount of light on a problem which was first seriously addressed by the civilisation before last and will certainly take at least a few more centuries to clarify to any great extent. Such sluggishness is the sign, in their eyes, that philosophy is dead.

That was the view taken by Stephen Hawking and his co-author Leonard Mlodinow in The Grand Design, at any rate. I have noticed as an empirical matter that when someone vigorously criticises philosophy, they are generally about to offer us some, and the rule does not fail in this instance: besides offering us some choice new a priori metaphysics, Hawking yields again to his predilection for putting Kant straight on a couple of points.

In fact, although the book gives a general potted history of physics (not bad apart from the ghastly jokes) , Hawking and Mlodinow are ultimately out to answer the fundamental questions of metaphysics: why is there anything? And why this?

The answer comes in two parts. There is something, because there's nothing to prevent a universe arising so long as certain requirements are balanced. Positive energy has to be balanced with negative energy; fortunately gravity provides negative energy of the kind that, for whole universes at a time, will serve to balance all the positive energy.

Why is it that there has to be this balance? Hawking and Mlodinow subscribe, it seems, to the old philosophical principle that nihilo ex nihil fit, or nothing will come of nothing, as King Lear put it. If things could appear out of nothing, then they might do so at any time or place, and the world would be incoherent: therefore, they can't. This principle has been part of the essential bootstrapping of many metaphysical theories although it's difficult to show convincingly why the world can't be incoherent (we just don't like the idea) and particularly difficult to show that it couldn't be just a little bit incoherent in certain cases and places and ways without having to descend into complete unintelligibility.

At any rate, the view offered here is that so long as the energies balance, the universe that springs into existence cancels out in theory and is therefore equivalent to nothing, so we're in the clear. It's a bit like pointing out that we can't create money out of nothing, but so long as there's a debt which matches the cash in our hands, the laws of the financial universe are satisfied. Handily it seems that the balancing of energies which allows whole universes to appear does not work within the universe, so that arbitrary entities cannot appear within the cosmos.

So far so good, if a bit skimpy; Hawking and Mlodinow don't give much attention to the question of whether there might be other constraints on the existence of universes (so that the balancing of energies might be a necessary but not sufficient condition of their existence); they seem to assume that there aren't. When we generate cash in exchange for a debt, we normally have a banker to satisfy, too; might not possible universes also face some additional hurdles before springing into existence? If not, don't we face the prospect of an incoherent series of slightly different universes, and is that really any different from a single universe incoherent in itself (in one case events are indeterminable because they're indeterminate in themselves; in the other they're indeterminable because you can't determine which universe you're in)?

They don't give much attention either to the question of different arrangements that might satisfy the balancing requirement. I got the impression that Hawking thinks something like our matter/energy entities and something very like gravity are the only real possibilities (rather in the way that you could incur debts in terms of cowrie shells or quatloos, but any medium of exchange is essentially money). There may be reasons for thinking this, but it would be good to know what they are. Is it a meaningless question to ask whether the cosmic balancing could be carried out in terms of say, 'left and right' or 'qwz and unqwz' rather than positive and negative? Perhaps, but then could a universe pull off a dual or triple constitution by achieving a balance of positive and negative values along two or three axes?

One reason Hawking and Mlodinow don't waste any time tidying up these loose ends is that they are relying on the second part of the argument, which explains why we've got this particular universe: the answer is the dreaded anthropic principle. The anthropic principle says the universe was bound to be one that was suitable for us to live in; there are strong and weak versions. Hawking and Mlodinow say the weak version dictates only our environment while the strong one governs the laws of nature too. I don't think that's quite right, although various statements of the difference have been offered. The weak version of the principle, as I understand it, is purely about appearances. It says that the world was bound to look to an observer like a place where observers could exist; but it's nothing to get excited about, any more than we should get excited about the lucky fluke that we were born on a planet that supports human life. The strong version, much more controversial, says that our existence has an actual causal or constitutive effect on the universe, and this is what Hawking and Mlodinow seems to be going for.

Reviews of the book generally highlighted the fact that Hawking had broken up explicitly with God. In the past he indulged the old fellow good-humouredly, rather as he might have done with a superannuated colleague, seeing no reason to brusquely attack his possibly-unjustified reputation and even speaking gravely of knowing his mind. Now, suddenly, he has no time for God.

The reason is actually quite clear: in order to make the anthropic principle seem plausible, Hawking and Mlodinow spend some time emphasising how exquisitely the fundamental constants and constitution of the universe are set up to create just those knife-edge conditions which make humanity possible; but that nice adjustment can be read another way. It's as though Hawking were making a speech about how no intelligent physicist can fail to be impressed by how exactly and non-randomly the universe has been designed for human beings; glancing at the audience he notices to his horror that the wrong people are nodding and hurriedly clarifies that the universe may be exquisitely designed, but for heaven's sake, not by God! Hawking and Mlodinow are a little sheepish about the nature of the anthropic principle:

this may sound like philosophy, they admit: I'm afraid it's rather worse than that - it sounds like theology.

They seek to defend the status of the principle as a scientific hypothesis, claiming that it leads to falsifiable predictions: for example, about the age of the universe. But doesn't seem to work; what we really do is deduce that the universe as we observe it could only be a certain age - the fact that we could only exist in a universe of this kind is another matter, established separately. From our mere existence we could not deduce the age of the universe at all, and the estimate of its age follows from observations to which our existence is actually irrelevant. But hey, that's OK - not everything has to be science.

Actually the case that Hawking and Mlodinow make for the precision engineering of the universe is not totally convincing either. Among other things they put forward the remarkable precision of the cosmological constant: but the cosmological constant is an arbitrary number chosen to make the sums come out right, with no other justification: it's there to fill a gap until a proper theory comes along. The only surprising thing about it is that physicists should be so impatient that they'd rather have an open lash-up like that than accept that for the time being that they don't understand the movement of galaxies. Impatience is similarly at work elsewhere; is it better to wonder at the inexplicable precision of theoretical constants, or hope that one day we might find an explanation for them? Would it have been better science if we'd rested on our laurels when the periodic table was established, contemplating in wonder how the elements had been arranged in such a neat way, sagely remarking that if they hadn't been arranged numerically we probably wouldn't be here, and that must surely be the reason for it?

One other problem with the strong form of the anthropic principle is that it requires that our present existence can reach back and influence the past of the universe. There is normally a strong presumption that the present cannot affect the past: if it does, then that past will change the present itself, and we get either a vicious circle or some kind of uncontrolled spiral and the world becomes incoherent again, because any event is subject to arbitrary revision (See how useful it is if the universe is not allowed to be incoherent!). Now Hawking and Mlodinow invoke the two-slit experiment (apparently it has now been performed successfully not just with photons, but with buckyballs, actual large molecules). You probably know about this famously perplexing business; for present purposes the key point is that if we look at where the particles are, their path changes. It looks as if our intervention now has somehow changed the direction they set off in just before: or if they're streaming in from a distant star, not just before, but long ages ago. Now I'm not sure that it's right to interpret the experiment as showing that we can change the past, but even if it is, there's a significant difference when it comes to influencing the constitution of the universe, of which the observer is inevitably a part: at the least it seems that in that case there's a particular problem of circularity.

There is, of course, another oddity about anthropicism: it seems to say that in the end the explanation for reality is to be found, not in the external world but in our own consciousness (phew - bet you thought I was never going to mention that). Now it's not the creationists in the audience who are nodding, but the idealists, the panpsychists, and perhaps even the solipsists; all rigorous thinkers but surely not the friends Hawking and Mlodinow were expecting for their proclaimed philosophy of 'model-dependent realism'?

Of course all this is intended to clear the way for M-theory. The kind of cosmic balancing Hawking and Mlodinow want from gravity requires supersymmetry and M-theory can provide it. Unfortunately M-theory, we're told, is not a grand unification of the kind we used to hope for: it

turns out those probably don't work. But so what, say Hawking and Mlodinow: after all, you can't have a map that shows the whole world (they seem to have an unusually jaundiced view of Mercator's projection), so why should we expect a single theory that accounts for everything? Instead we can have a family of different approaches to apply in different areas, just as we have different maps for different areas of the world. If we can't have a comprehensive final theory, let's take what we can have and proclaim that instead.

Well, the other approach might be to restrain our impatience and wait a bit to see what new insights come up. Science still seems to have a few problems to clear up; Hawking and Mlodinow mention a few of these, and they also describe Ptolemaic astronomy. The thing about Ptolemaic astronomy is that it actually worked rather well; if anything, the maths worked better than it did for the Copernican system, at least to begin with. The problem was that Ptolemaic astronomy was full of strange entities it was difficult to believe in and arbitrary values inserted simply to make the sums come out right. A change of paradigm was needed, and perhaps one or two small ones are needed now.

In fairness, I can understand the impatience for answers. It isn't necessarily a vice - and in a man like Hawking, who has spent so long with his apparent life-expectancy hovering only a little above zero, it is surely particularly understandable. But isn't there something depressing about the idea that philosophy is dead and science all but finished? Isn't there something more appealing in the idea that there's plenty more science to come yet, or as Newton put it:¶

> *I seem to have been only like a boy playing on the sea-shore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.*

God's Consciousness

9 February 2011

I worry sometimes about God. I suppose I've been a fairly consistent atheist all my life, though not particularly zealous: I realise that many people far cleverer than me have been committed believers in one religion or another and have written enlightening stuff about it which is well worth anyone's time to read. I don't really understand the apparently visceral hostility which people like Richard Dawkins display towards theism; for me it's more that in itself it doesn't seem a useful idea; instead it seems to be among the wilder pieces of metaphysics you could go in for (I enjoy wild metaphysics, of course - nothing better - but I don't adopt it). Be that as it may, I used to think that certain conceptions of God were rationally available if you wanted them. 'Geometrical' Gods would be one example, where you just define something like 'the universal total of awareness' or 'the origin of everything' as God, and then stick to it even though on careful examination your defined entity turns out to bear no apparent resemblance to the one people talk to in church. A little more scary, Gods that are absent, indifferent, unkind, or systematically deceptive seemed viable enough so long as you realise they're not really going to do any useful ontological work for you. They have a tendency to leave you with a theoretical mess slightly worse than the one you had to begin with, but if you want you can bring them along for the ride without egregious inconsistency.

Or so I thought: but in recent years I feel it's been getting more difficult to accommodate the idea of a non-material conscious entity. In the old days, when we had no idea how any of this worked, God was useful as the source of consciousness: he's got it and he bestows it on us. This view has not gone away and perhaps a modern version might be Peter Russell's equation of God and consciousness. If our consciousnesses are God it puts a strangely introverted complexion on prayer; but I think what he really means is that our consciousnesses are fragments of the universal version which is God. For me, that has too strong a flavour of gnosticism (roughly, the belief that we're all fragments of a divine being trapped by a ghastly cosmic accident in lumps of meat, and subject to a demiurge who falsely believes himself to be God. Why is gnosticism so popular, by the way? The secret esoteric doctrine always turns out to be gnostic. Come on, you heresiarchs: I was promised wild metaphysics!): I'm afraid for me gnosticism is one of those doctrines that resemble the legendary town in the mid-west of America that had a big sign saying Friend, if you've ended up here, you musta got on the wrong train somewhere.

At any rate, I find myself more in sympathy with Matt McCormick's case that an omnipresent God wouldn't be able to think because of his inability to draw the distinction between himself and the rest of the world. Although McCormick makes a cogent case, a determined theist could probably lash together a way of dealing with the problem of conscious omnipresence, but then there are others, possibly worse, associated with omniscience. God knows everything at once: he doesn't suffer from our form of the binding problem (the issue of how our brain makes a smooth coherent sequence out of sounds, vision, touch, etc) because all the correct things are linked up; but has a worse one of his own because everything is bound with everything else. He can't perceive some events as simultaneous and others as not, because he can't stop thinking

about any of them. How right the theologians were then, to say he lived in eternity rather than time: it was bound to look like eternity to him, at any rate, because he's incapable of perceiving change.

But surely that's all wrong: God can't have a binding problem because the binding problem relates to the synchronisation of sensory data; and as a non-material being, he can't have any senses (to see, you need material parts that can be affected by light, for example). But then he needs no senses, because he already knows where everything is and what it looks like. What he has a problem with is coherence and progression. Consciousness, we know, is a stream, and thought is a sequence, but it seems God is going to be incapable of either because everything is constantly in his mind.

But again, perhaps that to is wrong, and we need to distinguish between what God knows and what he's actually thinking about. We, after all, know lots of things without having them constantly in mind. That's an attractive way of looking at it because it suggests that although God knows about your sins, he will most likely never get round to giving them any attention, given all the things he has to think about - a comforting if heterodox view.

Picking on the 'omni' problems - omnipotence has its own long-standing difficulties - may be attacking a soft target; unfortunately there are less abstract ones, too. We sort of know by now that consciousness is associated with the activity of complex neural structures, and an immaterial God hasn't got neurons. We sort of know, too, that consciousness is associated with a family of cognitive abilities displayed in different degrees by different animals and arising out of a long process of evolution. We can see these abilites as being about the linking of sensory inputs and behavioural outputs, all the way from simple tropisms and reflexes through instinctive and thoughtful behaviour all the way to our largely detached meditations. We've already noted that God has no sensory inputs, and it's a little hard to see how a non-material being can have material outputs either. Our minds grew out of processes intimately connected with the needs of a physical animal - the 'three Fs' of feeding, fighting, and reproducing, all of which seem to be at best optional extras as far as God is concerned. Why would an immaterial being ever go through cognitive processes primarily designed to facilitate existence in a domain entirely alien to it, and if it did, what could it use to implement those processes?

Perhaps a theist would say that I'm harping far too much on the lower aspects of consciousness: sure, it helps us avoid the sabre-tooths and chip flints, and obviously God isn't much bothered in those areas; but what about the fancier aspects: what about qualia? A while back there did seem to be a move to recruit the Hard Problem as a crack in the otherwise solid wall of materialism which God might slip through, but it's hard to see how God could have phenomenal experience. To begin with, it seems he doesn't have sensory experience at all, for the reasons mentioned above; then again, if qualia are merely a matter of knowing what x is like he must have them in horrifying, mind-crushing, infinite simultaneous abundance.

What about intentionality, though? Perhaps God is a being composed of pure meaning, which would nicely account for his non-materiality. It looks attractive, but we surely don't want God locked up in the Platonic realm where could have only the status of an inert abstraction. Yet in the ordinary world intentionality only appears in two places: things made meaningful by us, and in those same old complex neural assemblies our brains, which God hasn't got.

These random noodlings of mine don't, of course, come near ruling out every possible conception of God (in fact the old man with a beard sitting on clouds starts to look pretty good,

so long as he’s not immaterial); but he seems to be slowly getting more and more problematic. In the past it’s always been possible to shrug and take it that God is so far above our comprehension we shouldn’t expect all the answers; but the conclusion that God can’t be conscious seems too alarming and at the same time too clear to be dismissed.'

Soul Dust

12 February 2011

Nicholas Humphrey is back with what he presents as a new attempt to knock that Hard Problem of qualia on the head once and for all. Soul Dust is not, however, a brand new theory: rather he has dusted down (or dusted up) the ideas he presented in Seeing Red, added some new points and a lot of persuasive advocacy, much of it in the form of quotes. He has many really good, interesting quotes, the sort that impel you to look up the original for more: but on the whole I think it was a mistake to invoke the 'raindrops on roses and whiskers on kittens' of The Sound of Music when extolling the splendiferous nature of a qualia-filled world.

The basic proposition of the book has two parts. The first is an account of perception and the self. Perception begins with a simple response to a stimulus - 'redding' if we're seeing red. There are no qualia involved in that, but when we form a loop by perceiving our own redding it adds a special vividness which we project back on to the perceived object; seen as a property of the object, this reverberation in our brain (Humphrey speaks of 'soul-hammering') naturally seems somewhat mysterious: it's like one of those paradoxical objects in an Escher picture (Humphrey uses as his example the 3D object made by Richard Gregory, which from one angle appears to be an impossible Penrose triangle). The looping also contributes to the construction of 'thick time', the perception of the present as something with significant duration rather than an infinitesimal line dividing past and present.

The enchanted quality of experience sheds the same illusory magic on the perceiving self, and by extension on the thinking Ego which goes with it. All this is why the world seems so worth experiencing and why we seem to ourselves so well worth preserving: a qualia-free zombie wouldn't have any great relish in life nor any particular reason to fear death. That's the second part of the case: this business of 'artificial' qualic impressions has come about through evolution because the extra zest it confers gives the possessor of qualia a survival edge over zombie counterparts.

The book is an engaging and persuasive read: you feel you've spent some time in friendly conversation with a very civilised and erudite man who has seen deeply into some of the issues. But there are a few problems.

The first is that the book isn't really about qualia at all. Humphrey, reasonably enough perhaps, dismisses the idea of ineffable qualia that can't be described and play no causal role in the world. Since he wants to give an evolutionary explanation, he has to take this stance: ineffable qualia couldn't have survival value. But if they ain't ineffable, they ain't really qualia, and what we're talking about ain't really the Hard Problem. It's the ineffability that's hard to explain; take that away and explaining why sensations are vivid may be non-trivial, but it's part of the Easy Problem. I'm not altogether sure he realises it, but Humphrey has pitched his camp with Dennett and accordingly all he is really entitled to say is "what qualia?".

Although they're not ineffable, Humphrey still wants his qualia to be illusions, a magic lantern show rather than a radar system. Why is that? If our perceptions are not of ineffable qualia, can't they just be of real qualities of actual things? One reason Humphrey wants them to be

illusions is that he wants to use the fact to validate that much-quoted phraseology about there being “something is it like” to experience things: the thing it’s like can be the thing itself. But also lurking somewhere in the back of Humphrey’s mind is, I suspect, that terrible old argument from error: because some of our perceptions are not of real things, none of them are - they’re all of proxies in our heads.

So everything that makes life seem worthwhile is actually an illusion? It’s a grim conclusion to reach, though Humphrey doesn’t seem to read it that way: he talks about the delight of experience and the wonder of the world as though he hadn’t begun by insisting that all these things are confidence tricks, quite false in fact. If Humphrey takes his own views seriously, he must be forced to conclude that in sober fact the zombies would be right: there actually is no particular reason to go on living and everything we value is in fact worthless. In fact, perhaps he does realise this, because he closes with an examination of the reasons why we go on living in the face of our undeniable mortality: not the actual reasons why we should go on living, but the deluded considerations that stop us topping ourselves immediately. He concludes that belief in an afterlife, supported by such things as our experience of waking every day, as if from death, is a natural part of the human outlook and the main thing that keeps most of us going. I think in fact that the majority of people, faced with the fact of eventual death, simply don’t waste time worrying about something they can’t change. Humphrey seems to think that a natural reaction to the inevitability of eventual death is immediate suicide: that might be one emotional reaction but I think to most people the lack of logic in bringing forward the very thing you fear is too salient to make that an attractive course. Myself I think it’s possible that if we attain sufficient maturity we might come to realise that one good life is enough. You don’t have to be a pessimist or severely depressed to think that rightly understood, the law that everything has its termination amounts to a promise, not a threat.

What about the second point, that qualia are here because they have positive survival value? I like the idea of looking for an evolutionary explanation, but I don’t think Humphrey ever makes this really convincing. The main argument seems to be that the extra pleasure given by qualia means we’re better motivated than zombies would be, and pay more attention. Yet there are surely many unpleasant and repulsive qualia; logically one might suppose as many of those as pleasant, attractive ones. Malthus tells us that most animals naturally live on the brink of starvation; having their constantly frustrated desire for food intensified doesn’t seem as obviously a good thing for them as it might be for us well-fed historical anomalies. Above all, aren’t qualia distracting? Isn’t it all too likely that while the qualophile animals are admiring the sunset they are more likely to be caught by the sabre-toothed tiger?

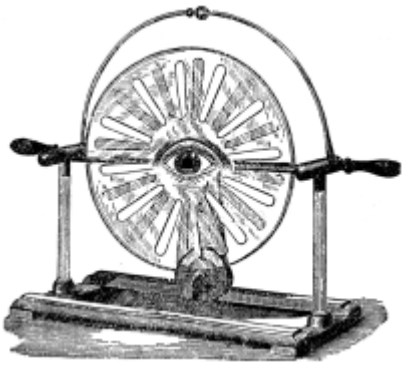
I think there is actually a potential argument here that Humphrey doesn’t use. Most animals get by very largely on instinct and other relatively hard-wired behaviour; only humans have really broken free to do whatever their rational deliberations - or their irrational whims - suggest to them. It could be argued that humans need qualia exactly to make the traditional sensory rewards for ‘good’ behaviour more potent in compensation for our greater ability to over-ride them. We need food to be more exciting and sex sexier to make us attend sufficiently to the basic biological imperatives instead of simply starving to death while reading or playing computer games (or committing suicide out of misplaced philosophical angst). That would make a certain kind of sense, anyway.

It may sound from the foregoing as if I radically disagree with everything Humphrey says, but that’s not the case at all: the book is a good rational effort in the right area, and it expresses

some complex insights with marvellous lucidity. But (alas, how many times have I said this over the years?) it's not quite the Answer.

Ephaptic Consciousness

21 February 2011



The way the brain works is more complex than we thought. That's a conclusion that several pieces of research over recent years have suggested for one reason or another: but some particularly interesting conclusions are reported in a paper in Nature Neuroscience (Anastassiou, Perin, Markram, and Koch). It has generally been the assumption that neurons are effectively isolated, interacting only at synapses: it was known that they could be influenced by each other's electric fields, but it was generally assumed that given the typically tiny fields involved, these effects could be disregarded. The only known exceptions were in certain cases where unusually large fields could induce 'ephaptic coupling ' interfering with the normal working of neurons and cause problems.

Given the microsopic sizes involved and the weakness of the fields, measuring the actual influence of these ephaptic effects is difficult, but for the series of experiments reported here a method was devised using up to twelve electrodes for a single neuron. It was found that extracellular fluctuation did produce effects within the neuron, at the miniscule level expected: however, although the effects were too small to produce any immediate additional action potentials, induced fluctuations in one neuron did influence neighbouring cells, producing a synchronisation of spike timing. In short, it turns out that neurons can influence each other and synchronise themselves through a mechanism completely independent of synapses'

So what? Well, first this may suggest that we have been missing an important part of the way the brain functions. That has obvious implications for brain simulations, and curiously enough, one of the names on the paper (he helped with the writing) is that of Henry Markram, leader of the most ambitious brain simulation project of all,Blue Brain. Things seem to have gone quiet on that project since completion of 'phase one'; I suppose it is awaiting either more funding or the advances in technology which Markram foresaw as the route to a total brain simulation. In the meantime it seems the new research shows that like all simulations to date Blue Brain was built on an incomplete picture, and as it stood was doomed to ultimate failure.

I suppose, in the second place, there may be implications for connectionism. I don't think neural networks are meant to be precise brain simulations, but the suggestion that a key mechanism has been missing from our understanding of the brain might at least suggest that a new line of research, building in an equivalent mechanism to connectionist systems could yield interesting results.

But third and most remarkable, this must give a big boost to those who have suggested that consciousness resides in the brain's electrical field: Sue Pockett, for one, but above all JohnJoe McFadden, who back in 2002 declared that the effects of the brain's endogenous electromagnetic fields deserved more attention. Citing earlier studies which had shown modulation of neuron firing by very weak fields, he concluded:

By whatever mechanism, it is clear that very weak em field fluctuations are capable of modulating neurone-firing patterns. These exogenous fields are weaker than the perturbations in the brain's endogenous em field that are induced

during normal neuronal activity. The conclusion is inescapable: the brain's endogenous em field must influence neuronal information processing in the brain.

We may still hold back from agreeing that consciousness is to be identified with an electromagnetic field, but he certainly seems to have been ahead of the game on this.

Mystery

7 March 2011


I'd like to say a word for mystery. I think Haldane summed it up:¶

> *…the universe is not only queerer than we suppose, but queerer than we can suppose.*


I hate that quote so much! The complacent fake modesty; the characteristic Oxford attitude of mingled superiority and second-ratism: don't you go thinking you can apply your mind to these weighty issues, little man; the best you can do is listen reverently to the words of our mighty predecessors. Footnotes to Plato! Footnotes to Plato!


Good grief, what a reaction! Well, then, let me quote someone you ought to like better; Leibniz:¶

…it must be confessed that perception and that which depends upon it are inexplicable on mechanical grounds, that is to say, by means of figures and motions. And supposing there were a machine, so constructed as to think, feel, and have perception, it might be conceived as increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, we should, on examining its interior, find only parts which push one another, and never anything by which to explain a perception.

¶It's interesting, incidentally, that Leibniz should have picked a mill. In these days of computers, it's natural for us to talk of thinking machines, but it must have been a much less obvious metaphor then; especially a mill, which doesn't produce any very complex behaviour… though Babbage called the central processor of the Analytical Engine the 'mill' didn't he… and of course Leibniz himself designed calculating machines, so perhaps a mechanical metaphor was more natural to him than it perhaps was to his readers… Anyway! The point is, this is a good example of a recurrent theme where someone holds up for our examination a mental phenomenon - in this case perception - and holds up as it were in the other hand the physical world, and says it's just obvious that the latter cannot account for the former.

Here's Brentano.¶

> *Every mental phenomenon is characterised by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction towards an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on. This intentional in-existence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it. We could, therefore, define mental phenomena by*

> *saying that they are those phenomena which contain an object intentionally within themselves*.

¶This time we're not talking about perception as such, but about intentionality: though Brentano claims it's characteristic of every mental phenomenon.

Then again, Nagel.¶

If we acknowledge that a physical theory of mind must account for the subjective character of experience, we must admit that no presently available conception gives us a clue how this could be done. The problem is unique. If mental processes are indeed physical processes, then there is something it is like, intrinsically, to undergo certain physical processes. What it is for such a thing to be the case remains a mystery.

It would be easy to find similar sources in which people contrast the pinkish-grey jelly in the skull with the sparkling abstract mental life it apparently sustains. These claims tend to have two things in common. The first is, they are essentially ostensive. There isn't really an argument being offered at this point, more a demonstration; we're just being shown. Look at this; then look at that; see?

Second, the claims are emphatic: they insist. It's obvious, they say, just look: no-one could think that this was that. These things have nothing whatever in common.


Of course they're emphatic: they want to hurry us on past the sketchy part before we pause and ask why this shouldn't be that. They've bundled the idea into its coat, thrust its hat into its hands, and are shoving it out the front door because they're afraid that otherwise we might entertain it for a while.


If they didn't want us to entertain the idea, why would they write about it? No. The absence of argument here isn't a weakness; on the contrary, it's a demonstration of the case. The very fact that we can't give arguments for the existence of our mental life proves its utter distinctness; we can't prove it, we can only notice it. But, and this is the reason for the emphasis, what noticing! 'In your face' doesn't begin to get it; phenomenal experience is inside your face; it's so emphatically there you could fairly say it defines the word.

What I'm saying is that these claims have a special quality; simply to pay careful attention to them is itself to experience their validity. The reality and distinctness of the mental world really deserve for once that much-misused term 'self-evident'.


It's going to be very convenient if the absence of argument is taken to make a claim self-evident. Proving six impossible things before breakfast will be child's play.

Actually, I think you are giving us an argument, but it's so feeble you prefer it not to be recognised: the argument from bogglement. It runs: I can't see how this could be that; it follows that it's somehow cosmically distinct. The falsity of the argument, once stated is too obvious too require further comment, but let me just remind you of all those people who couldn't imagine how the earth could possibly be moving.

Dennett has pointed out in a similar context that mysterians rely on passing off complexity as ineffability. Of course we don't know all the details of how particular physical events in the world trigger neuronal events in the brain, let alone how that vastly complex series of functions constitutes experience. Even if we had the information, we couldn't hold it all in mind at once. But our inability to do that, and any bogglement which may arise, does not in any way tend to show that there is no complete story.

Now Dennett would say that if only we could hold in mind all the details of the physical account, all this fantastically complex stuff, then the bogglement would vanish. But I'm not sure about that. Let me confess something: I too, boggle at the task of understanding the incomprehensible complexity of mental phenomena. But I boggle at other things too. Take computers. Now I think I can say without claiming to be an expert that I know how computers work. I've played around with one or two high-level programming languages; I've dabbled in machine code; I've run up a couple of routines for imaginary Turing machines. In short, I know how it works. And yet, it still sometimes looks like magic; my mind still boggles sometimes at what I see computers doing. Now if I can get bogglement from something I understand quite well and know is a 100% physical process, it follows that my boggling at the brain and what it does shows nothing.

 It's not the bogglement that matters; it's the undeniable direct experience. It's not because we don't understand experience that we say it's something distinctly non-physical; it's because we can see it's distinctly non-physical. To me I confess it seems to need some kind of educated perversity to deny this. Colin McGinn has pointed out that conscious experience is non-spatial: it has no position or volume. I don't know quite what we should say about that - abstractions like numbers are non-spatial too in what seems a different sort of way (if I mention Plato, will you start shouting again?); but it captures something about the obvious - patent - irreducibility of the mental.

I mean, just give it fair consideration; lift your eyes out of that two-dimensional world you're cycling round and round and just notice that there's another whole aspect to the world. To think it possible is in this case to realise it's true, I believe.

 You know, you're right.

 You see it?

 No, you're right that with you there is no argument.

Perplexities of Consciousness

16 March 2011

Eric Schwitzgebel, author of the excellent Splintered Mind blog (and Professor of Philosophy at University of California at Riverside) has a new book out, Perplexities of Consciousness, which sows doubt and confusion where they have never been sown before. Like Socrates, Schwitzgebel wants to make us wiser by showing us that we know much less than we thought. It has often been thought that while we might easily be wrong about the world and the things in it, we weren't prone to being wrong about how the world looks to us: Schwitzgebel seeks to show that in many respects we actually suffer from unresolvable confusions about even that, and worse, in some cases we're demonstrably wrong. Subjective experience may be a matter of there being something it is like to see red or whatever; but we actually have no clear idea of what that something is (or what it is like).

Back in 2007 Schwitzgebel published a book Describing Inner Experience? Proponent Meets Skeptic with Russell Hurlburt which examined Hurlburt's method of Descriptive Experience Sampling (DES). In DES subjects are asked to record their inward experience at random moments (prompted by a beeper) and are subsequently put through a fairly searching interview. This earlier book, as it happens, is currently the subject of a special issue of the Journal of Consciousness Studies. The book's title sets up a confrontation, but in fact it's clear that Hurlburt has a good deal in common with Schwitzgebel: the development of a special method for clarifying inner experience implicitly concedes that error is a serious possibility. The differences seem to be partly a matter of how well DES can really work, and partly a difference of agenda, with Schwitzgebel addressing the issues at a more radical and demanding philosophical level.

There is history to all this, of course: introspectionists like Wundt and Titchener once claimed that with careful training our inner experience could be described scientifically in great detail. The catastrophic collapse of that school of thought led on to the absurd over-reaction of behaviourism, which denied the very existence of inner experience. Only now, we might say, does it seem safe for Hurlburt to venture into the scorched territory of introspection and see if with a slightly different tack, a new beginning can be made.

Schwitzgebel's new book gives us plenty of intriguing reasons to be pessimistic about that project. The book shows its origins in a series of separate papers, but it does cohere around central themes, finding ambiguity and conflicting testimony just where we might hope for certainty.

First off it asks: do you dream in black and white? This used to be a common question and there was apparently an era when a large proportion of people thought they did. Nowadays no-one seems to think they dream without colour and in the more distant past the question never seemed to come up (dost thou dream in woodcut or tapestry?). Schwitzgebel very plausibly suggests that the idea of monochrome dreams is tied to the prevalence of black and white films and television. Perhaps if you asked, many people would now say they dream in 2D? I've had dreams that apparently took place at least partly in the world of some video game.

You may feel that the question is actually meaningless and that dreams don't typically either have or lack colour. Perhaps they are pure narrative, and asking whether they are in black and white makes no more sense than asking whether Pride and Prejudice was written in colour. Certainly if we extend the questioning and begin to ask whether dreams are in Technicolor, or what screen format or resolution they use (or in my case perhaps whether they were Xbox or PS2) the discussion starts to seem absurd; but could we deny that people might dream in black and white, at least sometimes? It doesn't seem too difficult to imagine that you might. It is possible that the prevalence of monochrome moving images actually changed the way people dreamed for a while; but it seems probable we must admit that some people were in fact mistaken about the qualities of their own dreams, and we must certainly accept that there was a degree of uncertainty. So Schwitzgebel has his wedge in position.

Second, he asks: do things look flat? He cites sources who say that a coin viewed at an angle looks elliptical: not to me, he objects (at least not unless I view it very obliquely, when I can sort of see it that way); if coins look elliptical to you does the world in general similarly look flat? Schwitzgebel addresses the idea that the coin in fact looks like its projection on to a flat surface and raises some objections: in fact, he claims there's no way to make the geometry of 'flattism' make sense. He suspects that here again people's intuitions have been captured by 'technology' - in this case people are thinking of how objects would be represented in a drawing or photograph. He remarks that some theorists have claimed that stereoscopic vision involves systematic doubling of perceived objects; while it's true that if we focus on something far away we can see two images of a finger held close to our face, Schwitzgebel finds it a very odd idea that most of the objects in our field of vision are normally doubled (I agree - I also agree with him that the world doesn't look all that much flatter when viewed with one eye rather than two).

Now we move on to academic, questionnaire-based research into our mental imagery, and it seems Galton is to blame for first spreading the idea that this sort of thing worked. Curiously, Galton's research found that the scientists in his sample were predisposed to deny the existence of mental imagery altogether, while the other subjects were more likely to accept it; a result which no-one has been able to duplicate since. Perhaps back then people thought mental imagery was an airy-fairy poetic business which hard-nosed scientists should reject. It turns out, moreover, that there is little or no correlation between reporting vivid mental imagery and being good at tasks which apparently require mental visualisation, such as comparing rotated 3d shapes. This is odd: why would evolution give us vivid mental imagery if it doesn't even help with tasks that require vivid mental images? Again it seems that our own reports are all but useless as a guide to what's really going on in there. We may claim to visualise a table, but under questioning we usually turn out to be unable to provide details, or become confused, or pause to imagine up some more details.

The next chapter raises the interesting question of human echolocation. Nagel famously took bats as his example of a creature whose inner experience we could not hope to imagine because it had a sense we entirely lacked. With amusing irony Schwitzgebel's case here is based on the fact that we do have some echolocation after all; we're just not normally aware of it. With a little practice we can learn how to stop short of a wall just by making regular noises and picking up the echo (if you're going to try this at home, I recommend taking some precautions to ensure that your nose isn't the first part of you to detect the wall unambiguously). Some blind people are well aware of this echolocating ability, but (a real score for Schwitzgebel) they misperceive it. Typically they describe the experience as being about

pressure on the face, and nothing to do with hearing: but experiments with blocked ears and covered faces show clearly that they're wrong about their own experiences.

I mentioned Titchener earlier: Schwitzgebel gives an interesting account of his methods. Whereas these days one would typically try to capture the subject's impressions as fresh as possible, uncontaminated by the experimenter's own prejudices, Titchener and his contemporaries took the opposite view: until you were very thoroughly trained in discrimination of your impressions, your testimony was worthless. Schwitzgebel explains how Titchener's researchers were trained to pick out 'difference tones', illusory notes heard under certain conditions. (You can try it out for yourself on Schwitzgebel's own page here. There is other useful stuff on his home page including abstracts and draft chapters.) Some of these tones are debatable and only a minority claim to be able to hear them: are the others failing to hear them, or hearing them and failing to discriminate? Titchener apparently has no answer.

The next chapter returns to earlier concerns about whether experience is sparse or abundant: are things outside the centre of our attention still constantly present in consciousness (abundant), or do they drop out (sparse). Schwitzgebel previously used the terminology 'rich' and 'thin', and we discussed some earlier Hurlburtian experiments of his. It will come as no surprise to find that Schwitzgebel, who quotes radically opposing views from a variety of sources, regards the matter as unresolved and possibly unresolvable; but I must say that this time round I couldn't see any reason why we shouldn't just conclude that experience is sometimes sparse and sometimes abundant. There's no doubt that sometimes when we focus narrowly on one thing, we lose track of everything else; and it seems hard to deny that at times we also pay vigilant attention to our surroundings in general. Couldn't it simply be that consciousness can have a wide or a narrow beam, at least partly under our own control?

Schwitzgebel now feels ready for a direct assault on the doctrine that our knowledge of our own experience is infallible. He warms up by questioning whether we know what emotion is, conceding (rather dangerously?) that his wife can sometimes judge his emotional state better than he can himself. What neutral yardstick he uses to confirm his wife's diagnosis is not altogether clear.

Why is it, he asks, that scarcely anyone, even the most vigorous sceptics, seriously questions the infallibility of introspection on certain points? The core argument seems to be that we can be wrong about the way things are, but we cannot be wrong about the way they appear to us. But why not? Schwitzgebel claims the argument rests on equivocation between two senses of 'appear' , one of them epistemic as in 'it appears to me that…'. I don't know whether the argument actually rests so much on the word 'appear' , but it seems a valid and interesting claim that there are two levels at work here: our experience and our beliefs or claims about it, with no special reason to think that the latter must be magically veridical.

Now there is a kind of rock-bottom argument available here; I don't think I've ever seen it used, but it may be in the back of the minds of those who argue for infallibility. This is that eventually you get below the level where truth and falsehood apply. If you pare experience down enough, you get to a point where it just is: it's not actually that it's infallible, more that it's beyond the realm of fallibility or infallibility. To be fallible, to have truth values, there has to be an element of intentionality (and the right kind, too, with an appropriate 'direction of fit' - I don't think desires can be false), but at the rock bottom level, this is absent. If I just experience without any thoughts about it at all, fallibility doesn't come into it.

I dare say Schwitzgebel might accept that up to a point, although he could reasonably point out that our experience is in practice completely suffused with a kind of intentionality; unconscious parts of our mind do an awful lot of interpretation before reports from our senses reach us and arguably pretty well everything is presented to us 'as' something, not just as mere sense-data (the kind of intentionality involved, that lets part of our brain add implicit messages to the conscious part about 'that there being a table' and so on is interesting, probably very important, and totally obscure); though generally it seems we can 'look through' to the basic sense-impressions if we want.

The question then is perhaps whether there is some very simple level of intentionality that can be added to the rock-bottom experience without any chance of its being wrong yet without it having the kind of trivial self-validation ('This is the sentence I wrote') that Schwitzgebel rightly excludes. Could it be that along with the experience itself we have an accompanying belief which says something like "Yeah, that…" which can't really be wrong? I still find it hard to resist the idea that there's something infallible in there.

It's a great point though; a strong and well-founded attack on such an important and well-accepted dogma has to be of great value. It forces us into greater clarity even if in the end it isn't accepted.

The book winds up with consideration of another engaging and interesting issue. What do you see when you turn out the light? (I can't tell you, but I know it's mine, as Lennon and McCartney argued in their seminal work.) Apparently little attention has been given to what things look like when our eyes are closed in normal light, but the book unearths quite a sequence of views about what we can see when our eyes are closed in the dark. Grey bands feature strongly in the older reports and then seem to drop out of favour: specks of light turn up fairly regularly, but you won't be surprised to hear that there is in the end no real consensus but a quantity of misplaced confidence. I think it's perhaps a little surprising so few people seem to say that when their eyes are closed in the dark they see nothing. Trying it myself I find I have to sort of insist on seeing and then get some very dim lines and grids, and flashes of amorphous shapes in brighter colour. This kind of thing might well have a good explanation in neurology and the more or less random firing of isolated neurons that respond to lines, grids, or patches of colour. At times, strangely, I had to do something hard to describe to stop my imagination intervening in what I was experiencing and livening things up. Overall, it doesn't look too promising a field for research to me, but, as Schwitzgebel suggests, why not see what you think (or see what you see)?

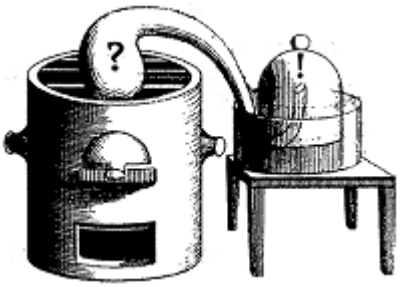
So in conclusion, what's the verdict? I think Schwitzgebel's case is essentially successful; his contention that there's more to deal with here than we may have realised seems hard to deny. This is salutary and also interesting; and since the subject is engaging and much of the discussion is readily accessible, I think the book deserves a wider readership outside the philosophical village.

I think it was Russell who said that when acting on a vigorous mind, scepticism produces new energy rather than despair: so can we add any positive conclusions to the overwhelmingly negative ones in the book? I'm left with two main thoughts. First, there is a philosphical case to be answered in the central assault on infallibility. Second, I think a great deal of the ambiguity and contradiction exposed by Schwitzgebel comes from the sheer complexity of the task. We ourselves, the experiencing entities, are pretty complex, with different levels of conscious and unconscious thought interacting in a variety of ways. Second, the ways we can experience and think about things is limitlessly varied. I don't think it would be difficult to list fifty significantly

different ways of ‘thinking about’ a table, with and without explicit imagery. Accordingly in many cases I think simple misunderstandings over what type of experience we’re talking about are more than half the problem. Perhaps Galton’s scientists thought ‘mental imagery’ meant what I would call ‘voluntary hallucination’; perhaps for those monochrome dreamers ‘black and white’ just meant dreams where colour wasn’t specifically salient. Taking all these problems together with a certain natural human variability, I think we might find explanations. It wouldn’t be trained subjects we need, just better terminology and a more clearly developed common understanding. I say ‘just’ - in fact it’s quite a tall order but not, I think, hopeless.

Experimental Free Will

27 March 2011



Shaun Nichols' recent paper in Science drew new attention to the ancient issue of free will and also to the somewhat trendy method known as 'experimental philosophy'. Experimental philosophy is liable - perhaps intended - to set the teeth of the older generation on edge for several reasons. One is that it sounds like an attempt to smuggle into philosophy stuff that shouldn't be there: if your conclusions can be tested experimentally they're science, not philosophy. We don't want real philosophy crowded out by half-baked science. It also sounds like excessive deference, as though some bullied geek started wearing football shirts and fawning on the oppressors. We may have to put up with the physicists taking our lunch money, but we don't have to pretend we like them.

Actually though, there doesn't seem to be any harm in experimental philosophy. All the philosophy that goes by the name appears to be real philosophy; the experiments are not used to clinch a solution but to help clarify and dramatise the problems. Often this works pretty well, and by tethering the discussion to the real world it may even help to prevent an excessive drift into abstract hair-splitting. Philosophers have always been happy to draw on the experiments of scientists as a jumping-off point for discussion, and there seems no special reason why they should do the same with experiments of their own.

In this particular case, Nichols shows that there is something odd about people's intuitive grasp of free will. Subjects were told to assume that determinism, the view that all events are dictated by the laws of physics, applied, and then asked whether someone would be responsible for various things. In the vaguest case they all agreed that in general, given determinism, people were not responsible for events. Given a specific example of a morally debatable act they were less sure; and when they were offered the example of a man who takes out a murder contract on his wife and children, most felt sure he was responsible even given determinism.

This is odd because it's normally assumed that determinism means no-one can be responsible for anything. In order to be responsible, you have to have been able to do something else, and according to determinism the laws of physics say you couldn't have done anything but what you did. It's odder because of the distinction drawn between the cases.

It could be that something in the experiment predisposed subjects to think they were required to make distinctions of this kind, or it could be that ordinary subjects are not very good at coming up with strictly logical consequences of artificial assumptions; but I don't think that's really it. The distinction between the three cases appears to be one of blame -so it looks as if the man in the street doesn't really grasp the philosophical concept of responsibility and relies instead on a primitive conception of blameworthiness!!!

But, um - what *is* the philosophical concept of responsibility? It's pretty clear when we cite the laws of physics that we're talking about causal responsibility - but causal responsibility and moral responsibility don't coincide. It's clear that you can be causally responsible for an event without being morally responsible: someone pushed you from behind so that you in turn pushed someone under a train. Less clearly, in some cases it is held that you can be morally responsible for events you didn't knowingly bring about: the legal doctrine strict liability, Oedipus bringing a

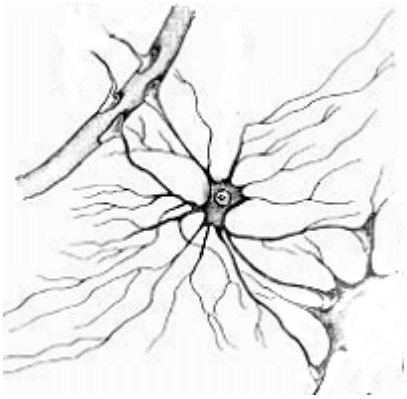
curse on Thebes, poor Clarissa wondering whether having been raped is in itself a sin. All of these are debatable; we might be inclined to see strict liability as a case of legal overkill: “we care so much about this that we’re not even going to entertain any discussion of responsibility - so you’d better make damn sure things are OK” . In the other cases we typically think the assignments of blame are just wrong (although Milan Kundera notably reclaimed the moral superiority of Oedipus in The Unbearable Lightness of Being). But the distinction between moral and causal is clear: does the determinist case equivocate between the two, and were Nichols’ subject too shrewd to be taken in?

If moral responsibility and the assignment of blame is a social, conventional matter, then it seems it might be so. No-one would suggest to a writer that he was not the author of his novel because it was all the result of the laws of physics, although in one sense it’s so. No-one would accept on similar grounds that I’m not responsible for a debt, however abstract and conventional the notions of debt and money may be compared with the rigorous physical account of events. So why should should the physical story stop us concluding that on another level of description we can be interestingly and coherently blameworthy? That would be a form of compatibilism, the view that we can have our determinist cake and eat our free will, too. So perhaps Nichols’ subjects were compatibilists.

That would be an interesting discovery but… an interesting discovery in psychology. The fact that most people are instinctively compatibilists provides no particular reason to think compatibilism is true. For that, we still have to do the philosophy the old-fashioned way. Scientists may be able to gather truth from the world, like bees with nectar: philosophers are still obliged, like bees, to spin their webs out of their own internal resources.

State Space and the Triumph of the Astrocytes

9 April 2011



Alfredo Perreira Jnr has kindly let me see an ambitious paper he and Leonardo Ferreira Almada have produced: Conceptual Spaces and Consciousness: Integrating Cognitive and Affective Processes (forthcoming in the International Journal of Machine Consciousness).

The unifying theme of the paper is indeed the integration of emotional and neutral cognitive processes, but it falls into two distinct parts.

The first, drawing on the work of Peter Gärdenfors, sets out the heady vision of a universal state space of consciousness. Such a state space, as I understand it, would be an imaginary space constructed from a large number of dimensions each corresponding to one of the continuously variable aspects of consciousness. In principle this would provide a model of all possible states of consciousness, such that anyone's life experience would form a path through some area of the space.

The challenges involved in actually populating such a theoretical construct with real data are naturally daunting. Perreira and Almada suggest that it could be approached on the basis of reported states of consciousness. An immediate problem is that qualia, the essence of subjective experience, are widely considered to be unreportable: Perreira and Almada meet this head on by adopting the heterophenomenology of Daniel Dennett: this approach (which I think implies scepticism about ineffable qualia) is based on studying phenomenal experience indirectly, through what subjects say about their own: the third-person perspective. Perreira and Almada note that Dennett adopted this stance mainly as a means of refuting first-person approaches, but I'm sure he would (or will) be delighted to hear of its being adopted as the explicit basis of serious research. It's implicit in this approach that we're dealing with states that are capable of 'inter-subjective validation', that is, that they're states which are accessible to all conscious entities. This rules out objections on the grounds that, say, Andy having experience X is not the same as Bill having experience X, though in so doing it may appreciably impoverish the scope of the exercise. It could be that the set of experiences common to all conscious beings is actually a significantly restricted sub-set of the whole realm of conscious experience. For that matter, can we afford to ignore the unconscious or the subconscious? At times the borderline between conscious states and their near relations may be blurry.

I think two other worries are worth a mention. The state space model suggests that all trajectories are equally valid, but it seems unlikely that this is the case here. Consciousness is a stream, both emotionally and cognitively: certain kinds of state naturally follow other kinds of state. In fact, it doesn't seem too much to claim that some states refer to previous states: we can't repent our anger intelligibly without having first been angry. The business of reference, moreover, is a problem in itself. We've already excluded the possibility of Andy's anger being different from Bill's, but we can also be in the same state of anger about different things, which seems a material difference. Me being angry about my tax return is not really the same state of consciousness as me being angry about receiving a parking ticket, though in principle the anger itself could be identical. Because, thanks to the miracle of intentionality, we can be angry about anything, including imaginary and logically absurd entities, this is a large problem. Either we

exclude these intentional factors - and put up with a further substantial impoverishment of our state space - or the size of our state space balloons out infinitely in all directions.

The practical problems are not necessarily fatal, of course: it's not as if Perreira and Almada were actually proposing a fully-documented description of the universal state space. What they do suggest is that if we assume another state space (wow!) corresponding to all the possible biophysical states of the brain, we can then hypothesise a mapping of points in one space to points in the other, which would give us the prize of a reduction of conscious experience to physical brain function. Now I think a biophysical state space of the brain faces formidable difficulties of its own: for one thing we really don't know exactly which biophysical features of the brain are functionally relevant; for another different brains are not wired the same way - and of course the sheer complexity of the thing is mind-boggling. The biophysical state space of a single neuron is a non-trivial proposition.

However, at a purely theoretical level, this is a nice rigorous statement of what the much-sought Neural Correlates of Consciousness might actually be. If we merely claim that there is a mapping between the two state spaces, we have a sort of rock-bottom version of NCCs, a possible statement of the minimum claim. We would expect there to be some more general correspondences and matches between regions and trajectories in the two spaces - though I think it would be optimistic to expect these to be simple (and constructing the two state spaces and then observing the regularities would be a remarkably long way round to discovering correspondences between brain and mind activity). Still, the fact that these pesky NCCs turn out to be more abstract and problematic than we might have hoped is in itself a conclusion worthy of note.

All these heroic speculations are in any case just the hors d'oeuvres for Perreira and Almada: the state space of consciousness would have to represent emotional affect as well as rational cognition: how would that work? They proceed to review a series of proposals for integration emotion and cognition. Damasio's Somatic Marker Hypothesis, which has emotional affect deriving from states of the body is favourably considered, though criticised for elements of circularity. The alternative view that emotions come first avoids such problems but is ciriticised for not squaring with empirical evidence. Perreira and Almada suggest a better third alternative might be based on mapping the complex inter-relations of cognition and effect, and give a friendly mention to the oft-quoted Global Workspace theory of Bernard Baars. Now we can begin to see where the discussion is going, but first the paper brings in a new element.

This is a discussion of what actually makes mental states conscious - embodiment, higher order states, or what? Perreira and Almada look at a number of proposals, including Arnold Trehub's retinoid system and Tononi's concept of Phi, a measure of integrated information. Briefly, they conclude that something beyond all these approaches is needed, and something which puts the integration of affect and cognition at the heart of the system.

Now we come to the second major part of the paper, where Perreira and Almada introduce their own proposal: step forward the astrocytes.

Astrocytes are the most common form of glial cell, which are 'the other brain cells'. Neurons have generally had all the glory in the past; historically it was assumed that the role of glia was essentially as packing for the neurons - in fact 'glia' is the Greek word for 'glue'. In recent years, however, it has become gradually clearer that glia, and astrocytes in particular, are more important than that. They form a second network of their own, across which 'calcium waves' are

propagated. I think it would be true to say that the standard view now sees astrocytes as important in supporting and modulating neural function, while still reserving the main functional signficance for all those showy synaptic fireworks that neurons engage in. Perreira and Almada want to give the neural and glial networks something like parity of esteem. The proposal, in essence, is that plain cognition is done by the neurons, while feelings are carried by large astrocytic calcium waves: only when the two come together does consciousness arise. Consciousness is the "astroglial integration of information contents carried by neurons".

What about that? It's a bold and novel hypothesis (something we certainly need); it's at least superficially plausible and has a definite intuitive appeal. But there are objections. First, there seem to be other established candidates for the role of feeling-provider. We know that certain parts of the brain are required for certain kinds of affect - the role of the amygdala in producing 'fear and loathing' (or perhaps we should say 'reasonable distrust and aversion') has been much discussed. Certain emotions are almost proverbially (which of course is not to say accurately) associated with hormones and the action of certain glands. This needs to be addressed, but I don't think Perreira and Almada would have too much difficulty in setting out a picture which accomodated these other systems.

More difficult I think are two more fundamental questions. Why would astrocytic calcium waves cause, or amount to, feelings? And why would those feelings, when associated with cognitive information, constitute consciousness? Damasio's and other theories can offer a clearer answer on the first point because it's at least plausible that emotional states can be reduced to the pounding heart, the watering eyes, the churning stomach: calcium waves rippling across the brain somehow don't seem as obviously relevant. And then, is it really the case that all conscious states have emotional affect? Perreira and Almada suggest that if neurons alone are involved (or astrocytes alone) all you get is proto-consciousness: but intuitively there doesn't seem anything difficult about completely dispassionate but fully conscious thought.

One strength of the theory is that it seems likely to be more open to direct scientific testing than most theories of consciousness: a few solid experiments would probably relegate the kind of objection I've mentioned to secondary status. So perhaps we'll see...

Eagleman's Law

17 April 2011

I thought David Eagleman's book SUM was excellent - it's a series of short accounts of the afterlife in which each version turns out to be surprising and disappointing in a variety of imaginative ways. So I looked forward to Incognito, his serious account of the conscious mind. The dazzling striped lettering on the dustjacket suggests we're in for a lively read, and it does not mislead.

The first part of the book - the bulk of it in fact - is a highly readable account of many of the interesting and counter-intuitive discoveries of modern research about how the brain in general, and perception in particular, works. There are many old friends here - blindsight, split brains, and synaesthesia, and so on - but also stuff that was new to me. If you want to know what Japanese chicken-sexers and WWII British plane spotters have in common, this is the book for you. In places I felt Eagleman's account could have benefited from a few caveats and qualifications, and at times the breeziness of his explanation seems to carry him away. Is the tendency to talk freely about your life to anonymous strangers, people you might meet on a train journey, for example, really the 'explanation' for the institution of confession in the Catholic church? Well, no. But in fairness this is a popular account, not a scientific paper.

As the book progresses it becomes clear that Eagleman does have some points of his own to make. He speaks warmly of Minsky's Society of Mind, but suggests that to complete the picture we need to assume that there is an ongoing competition for control among the various agents running our minds. I don't think this idea is quite as novel as Eagleman seems to suppose, and he goes on to make a very traditional use of it by drawing a distinction between a rational controller, able to defer gratification, and a short-term pleasure seeker. This sort of echoes Freud, and indeed Plato's charioteer.

In fact Eagleman's main purpose is to change our view of moral responsibility and legal responses to crime. He spends some time quoting examples of people who committed crimes under the influence of drugs or brain tumours, and recounts the well-known story of Phineas Gage, whose behaviour was changed for the worse when a tamping iron was accidentally fired through his brain. This last example perhaps needs to be handled with a little more care than Eagleman gives it, as there are reasons for a degree of scepticism about how changed or how bad Gage's behaviour became. Eagleman foresees a day when neuroscience will make it impossible to hang on to the idea that free will means anything or that anyone is ultimately responsible for anything - or as puts it blameworthy.

Eagleman seems to have a bit of a campaign going against blame. His base is at Baylor College of Medicine in Houston where he directs the College's Initiative on Neuroscience and Law (as well as the Eagleman Laboratory for Perception and Action). He approvingly quotes Lord Bingham:

> *In the past, the law has tended to base its approach... on a series of rather crude working assumptions: adults of competent mental capacity are free to choose whether they will act in one way or another; they are presumed to act rationally, and in what they conceive to be their own best interests; they are credited with such foresight of the*

> *consequences of their actions as reasonable people in their position could ordinarily be expected to have; they are generally taken to mean what they say. Whatever the merits or demerits of working assumptions such as these in the ordinary range of cases, it is evident that they do not provide a uniformly accurate guide to human behaviour.*

At first sight it's possible to read this as a liberal, even humane point of view: isn't it ridiculous the way we blame and punish all these people for things they couldn't help? But a moment's reflection shows that the noble lord is letting these common folk off their punishment only because he considers them sub-human. You know, he says, for some reason our legal system bends over backwards to treat people as if they were dignified creatures of some moral standing: it goes on treating them as worthy of admonition, worthy indeed of the honour of punishment, long after they have proved themselves to be pond life. You would almost think the law considered these people as in some sense the equal of their judges!

The more I read the quote the less I like it.

Eagleman certainly has no truck with equality and devotes a section to denouncing it. The question of whether equality of income is in itself a moral desideratum or an economic advantage, or the reverse, is of course politically contentious: but I had thought that no-one anywhere on the spectrum denied moral equality. Can Eagleman possibly mean that we don't all equally deserve a fair hearing, natural justice, and to be judged by our actions alone? Reader, I fear he does.

In fact what Eagleman is advocating in practice is not all that clear to me in any detail, but a couple of points are well established. People will still go to jail, he is keen to emphasise: not in order to be punished, but in order to protect society. It follows, though Eagleman does not dwell on this, that they might stay in jail forever, if they continue to be deemed dangerous. Indeed, if neurolaw gives them the thumbs down as high-risk future perps, it seems to follow that they might find themselves locked up before they've done anything.

The other point, a little more appealing, is that Eagleman believes people can be trained out of their vicious propensities. Crime, he suggests, occurs when the struggle between the different agents in our mind goes the wrong way: when the rational self loses out to the weak-willed glutton. He mentions experiments conducted by his colleagues, which use neural feedback to try to help people overcome their desire for chocolate cake: he thinks a similar 'prefrontal workout' might help criminals get ready to re-enter society. So perhaps they won't be in jail long, after all.

But really, does crime arise from weakness of will? Is it that everyone wants to be good, but sometimes some people give way to an immediate temptation? Isn't that a rather minor part of the problem? I can't help thinking what a sunlit, untroubled life Eagleman must have led, never having experienced in himself or having reason to suspect in others any of the deep dark recesses of the soul, if he thinks evil is more or less the same as finding it hard to give up smoking.

Another instance of naivety seems embodied in the idea that in the pandemonic struggle of our minds one side is always good and the other bad. Suppose we train our criminals to overcome their temptations and enthrone the rational, long-term part of their brain - will that make them model citizens, or psychopaths who've learnt to rein in the empathy and repugnance which would otherwise have prevented their crimes? Terrible things have been done on grounds that seemed entirely cold and rational to their perpetrators. Sometimes it's better to leave Falstaff in

charge: there may be a glut of chocolate cake and feasting, but the battle of Shrewbury will be cancelled.

Eagleman is aware of the poor precedents for science-driven justice, but he has a curious way of immunising his own mind against them. He describes the failure of psychologists to predict the rate of re-offending: he describes lobotomy with distaste, yet somehow manages to construe them as relics of the traditional way of doing things rather than early attempts at the kind of science-driven crime management he too is advocating (of course, they got it wrong while he will get it right). He describes the plot of One Flew Over the Cuckoo's Nest without seeming to realise how much the predicament of McMurphy resembles that of an inmate in a new Eaglemanic detention facility. When McMurphy was in jail, under the benighted old system, his punishment was limited to match his crime and when his term was up a system like that deplored by Lord Bingham persisted in trying to make a free moral agent of him again. In the asylum, as in Eagleman's jail, you don't get out till the men in white coats are pleased with you.

At the end of the day, there's a fundamental concept missing from Eagleman's analysis: Justice. He's not the only one in recent times to overlook it or mistake it for revenge, but I'll go out on a limb and say that it's fundamental. There is a moral imperative that we should ensure good behaviour leads to good things and bad to bad regardless of other considerations, and this is what lies at the root of all punishment. Deterrence and rehabilitation are additional benefits we should strive to secure, but justice is what it's about. Only justice authorises, in theory and in practice, the exercise of governmental or judicial power, and to put it aside, however good the reason, is tyranny.

Coming off the high horse, I suppose we can thank Eagleman for stating explicitly conclusions which others have avoided or fudged, and thereby promoting the clarity of debate. He could go a bit further in setting out the implications of his theses and proposing what they might really mean in practice.There's no doubting that our current judicial systems are far from perfect, and even if we reject Eagleman's prescriptions, they might in the end help things move on.

Political Consciousness

7 May 2011

As Gilbert and Sullivan had it,¶

> *…every boy and every girl*
> *That's born into the world alive*
> *Is either a little Liberal*
> *Or else a little Conservative!*

Now indeed it seems that right-wing brains are measurably different from left-wing ones.

Well, all right, it's not as simple as that, but as a review of an interesting piece of research reports, in one sample of young adults, self-reported political conservatives tended to have a larger right amygdala, whereas self-reported liberals tend to have a larger anterior cingulate cortex. Strictly speaking we are probably not entitled to deduce anything from this, but if we want to jump to conclusions we can assume that a larger amygdala is associated with greater levels of distrust and hostility, while a larger anterior cingulate cortex is associated with greater tolerance of conflict and uncertainty. It's easy to imagine that a more distrustful (perhaps we should say 'sceptical') personality, with a more pessimistic view of human nature, might be associated with generally more right-wing views. I'm not so sure about the interpretation of the other finding, but I suppose a greater tolerance of conflict and uncertainty might be associated with less support for established authority and traditional mores, and so with a generally more leftish slant. I suppose we would expect, in line with the often-observed tendency to drift to the right as one ages, that the amgydala would swell and the anterior cingulate cortex shrink as time went by?

It's generally fairly plausible that political views correlate to at least some degree with personality traits. It's often suggested that introversion tends to go with a more conservative outlook, for example. It's certainly the case that political leanings are more about direction of travel than about specific policies: pretty well everyone alive now is a raving lefty in terms of the medieval political outlook (though even then there were outbursts of radicalism, of course). Conservative opinion which once would fight to the death defending the divine right of kings finds itself, four hundred years later, defending the mercantile individualist values it once regarded as the enemy. For that matter I've just been reading a few of Trollope's novels, and it seems pretty clear that the Duke of Omnium's coalition government would find its successor in Westminster today not just unacceptably but almost incomprehensibly left-wing, even though in theory the governments have a broadly similar mixture of conservative and liberal opinion.

It's strange that Kanai's research identifies two different measures of political leaning. It doesn't seem to me that trusting and liking people is exactly the same thing as tolerating conflict. The existence of two independent axes suggests that there are actually four different positions. As well as the die-hard right and the entrenched left, we have on the one hand people who favour hard-nosed policies but attach no value to authority and convention; while on the other hand we have people who prefer generous and supportive systems but want to combine them with traditional and conforming ways of behaving. Neither seems at all hard to imagine: I know people who would fit both those descriptions tolerably well. Perhaps our existing political set-up misses out on the full geometry: if so it might be a little worrying because it would imply that

perhaps there are problems and solutions which are really quite important but remain partly invisible to us on our one-dimensional view?

If the normal right-left spectrum is so inadequate (and I think many people would say it is, and that there are really not just two axes involved, but several), how come we're lumbered with it? It's not that surprising that when one group is in power many of its opponents band together and sink their differences in the joint project of turning the rascals out; and when the tables are turned it's the new opposition that benefits from the same effect. Over time it seems plausible this would lead to the crystallisation of two broad groupings; and the theme of those who have established wealth and power, and hence tend to like theoretical arguments against change, versus those who have less and so are predisposed to like reforming and revolutionary sentiments, seems well adapted to be the thread along which the parties ultimately take shape.

Fine for politics, but I can't see how that would explain a similar dichotomy in the brain, unless there's something more fundamental going on. Could there be some kind of game-theoretical account, in which, say, big-amygdala people do well out of their hard-nosed attitudes and reproduce successfully up to the point where they become a majority of the population, at which point the level of distrust undermines social cohesion and fully offsets the benefits to new 'selfish' entrants, so the advantage begins to accrue to the cingulate cortexers who can cope with a bit of dissent and disorder, who then do better until their generosity and trust begins to be exploited by a selfish minority and so on until some kind of balance is reached?

Maybe it's not genetic, though: perhaps it has to do with birth order? It is said that first-born children tend to endorse authority and take on the role of parental deputy, while for subsequent children there are more rewards in being the first non-conformist than in being a mere second deputy. Perhaps these influences create differential growth in differing parts of the brain (alas, no data on birth order)?

Having jumped repeatedly from one conclusion to another like an excitable frog, I find myself in an unfamiliar part of the pond, so I shall retreat, taking with me only the wild speculation that much of the theory and rhetoric of politics might in fact resemble the bizarre confabulations produced by some patients to give a superficial appearance of conscious volition to behaviour whose actual origins in deep mental hard-wiring or cognitive deficits are actually quite unknown to them.

Simples Consciousness

15 May 2011

There is a lot of interesting stuff over at the The Third Annual Online Consciousness Conference; I particularly enjoyed Paul Churchland's paper Consciousness and the Introspection of Apparent Qualitative Simples (pdf), which is actually a fairly general attack on the proponents of qualia, the irreducibly subjective bits of experience. Churchland is of course among the most prominent, long-standing and robust of the sceptics; and it seems to me his scepticism is particularly pure in the sense that he asks us to sign up to very little beyond faith in science and distrust of anything said to be beyond its reach. He says here that in the past his arguments have been based on three main lines of attack: the conditions actually required for a reduction of qualia; the actual successes of science in explaining sensory experience, and the history of science and the lessons to be drawn from it. Some of those arguments are unavoidably technical to some degree; this time he's going for a more accessible approach and, as it were, coming for the qualophiles on their own ground.

The attack has two main thrusts. The first is against Nagel, who in his celebrated paper What is it like to be a bat? claimed that it was pointless to ask for an objective account of matters that were quintessentially subjective. Well, to begin with, says Churchland, it's not the case that we're dealing with two distinct realms here: objective and subjective overlap quite a bit. Your subjective inner feelings give you objective information about where your body is, how it's moving, how full your stomach is, and so on. You can even get information about the exhausted state of certain neurons in your visual cortex by seeing the floaty after-image of something you've been staring at. Now that in itself doesn't refute the qualophiles' claim, because they go on to say that nevertheless, the subjective sensations themselves are unknowable by others. But that's just nonsense. Is the fact that someone else feels hungry unknowable to me? Hardly: I know lots of things about other people's feelings: my everyday life involves frequent consideration of such matters. I may not know these things the way the people themselves know them, but the idea that there's some secret garden of other people's subjectivity which I can never enter is patently untrue.

I think Churchland's aim is perhaps slightly off there: qualophiles would concede that we can have third-person knowledge of these matters: but in our own experience, they would say, we can see there's something over and above the objective element, and we can't know that bit of other people's feelings: for all we'll ever know, the subjective feelings that go along with feeling hungry for them might be quite different from the ones we have.

But Churchland has not overlooked this and addresses it by moving on to the bat thought-experiment itself. Nagel claims we can't know how it feels to be a bat, he says, but this is because we don't have a bat's history. Nagel is suggesting that if we have all the theoretical information about bat sensitivity we should know what being a bat is like: but these are distinct forms of knowledge, and there's no reason why the possession of one should convey the other. What we lack is not access to a particular domain of knowledge, but the ability to have been a bat. The same unjustified claim that theoretical knowledge should constitute subjective knowledge is at the root of Jackson's celebrated argument about Mary the colour scientist, says

Churchland: in fact we can see this in the way Jackson equivocates between two senses of the word 'know': knowing a body of scientific fact, and 'knowing how' to tell red from green.

The second line of attack is directed against Chalmers, and it's here that the simples of the title come in. Chalmers, says Churchland, claims that a reductive explanation of qualia is impossible because subjective sensations are ultimately simples - unanalysable things which offer no foothold to an inter-theoretical reduction. The idea here is that in other cases we reduce away the idea of, say, temperature by analysing its properties in terms of a different theoretical realm, that of the motion of molecules. But we can't do that for subjective qualities. Our actual experiences may consist of complex combinations, but when we boil it down enough we come to basic elements like red. What can we say about red that we might be able to explain in terms of say neurons? What properties does red have? Well, redness, sort of. What can we say about it? It's red.

Churchland begins by pointing out that our experiences may turn out to be more analysable than we realise. Our first taste of strawberry ice cream may seem like a simple, elemental thing, but later on we may learn to analyse it in terms of strawberry flavour, creaminess, sweetness, and so on. This in itself does not prove that there isn't a final vocabulary of simples lurking at the bottom, of course. But, asks Churchland, how will I know when I've hit bottom? Since every creature's ability to discriminate is necessarily limited, it's inevitable that at some point it's going to seem as if I have gone as far as I could possibly go - but so what? Even temperature probably seemed like a simple unanalysable property once upon a time.

Moreover, aren't these unanalysable properties going to be a bit difficult to handle? How do we ever relate them to each other or even talk about them? Of course, the fact that qualia have no causal properties makes this pretty difficult already. If they don't have any causal effects, how can they explain anything? Qualophiles say they explain our conscious experience, but to do that they'd need to be registered or apprehended or whatever, and how can that happen if they never cause anything? As an explanation, this is 'a train wreck'.

Churchland is quite right that all this is a horrible mess, and if Chalmers were offering it as a theory it would be fatally damaged. But we have to remember that Chalmers is really offering us a problem: and this is generally true of the qualophiles. Yes, they might say, all this stuff is impossible to make sense of; it is a train wreck, but you know, what can we do because there they are, those qualia, right in front of your nose. It's pretty bad to put forward an unresolved mystery, but it would be worse to deny one that's palpably there.

On the point about simples, Churchland has a point too: but there does seem to be something peculiarly ungraspable here. Qualia seem to be a particular case of the perpetual give-away argument; whatever happens in the discussion someone will always say 'the trouble is, I can imagine all that being true, and yet I can still reasonably ask: is that person really having the same experience as me?' So we might grant that in future Churchland will succeed in analysing experience is such a way that he'll be able to tell from a brain scan what someone is experiencing, conclusions that they will confirm in great detail: we can give him all that and still feel we don't know whether what we actually experience as red is what the subject experiences as blue.

Churchland thinks that part of the reason we continue to feel like this is that we don't appreciate just how good some of the scientific explanations are already, let alone how good they may become. To dramatise this he refers back to his earlier paper on Chimerical colours . It

turns out that the ‘colour spindle’ which represents all possible colours is dealt with in the brain by a neuronal area which follows the Hurvich-Jameson model. The interesting thing about this is that the H-J model is larger than the spindle: so the model actually encodes many impossible colours, such as a yellow as dark as black. Presumably if we stimulated these regions with electrodes, we should experience these impossible colours.

But wait! There is a way to hit these regions without surgery, by selectively exhausting some neurons and then superimposing the afterimage on a coloured area. See the paper for an explanation and also a whole series of practical examples where, with a bit of staring, you can experience colours not in nature.

These are well worth trying, although to be honest I’m not absolutely sure whether the very vivid results seem to me to fall outside the colour spindle: I think Churchland would say I’m allowing my brain to impose sceptical filtering - because some mental agent in the colour processing centre of my brain doesn’t believe in dark yellow, for example, he’s whispering in my ear that hey, it’s really only sort of brown, isn’t it?

For Churchland these experiments show that proper science can make predictions about our inner experience that are both remarkable and counter-intuitive, but which are triumphantly borne out by experience. I do find it impossible not to sympathise with what he says. But I can also imagine a qualophile pointing out that the colour spindle was supposed to be a logically complete analysis of colour in terms of three variables: so we might argue that these chimerical colours are evidence that analyses of sensory experience and the reductions that flow from them tend to fail and that the realm of colour qualia, quite contrary to the apparently successful reduction embodied in the colour spindle, is actually unconstrained and undefinable. And why are these experiments so exciting, the qualophile might ask, if not because they seem to hold out the promise of new qualia?

Beyond Libet

31 May 2011

Libet's famous experiments are among the most interesting and challenging in neuroscience; now they've been taken further. A paper by Fried, Mukamel and Kreiman in Neuron (with a very useful overview by Patrick Haggard) reports on experiments using a number of epilepsy patients where it was ethically possible to implant electrodes and hence to read off the activity of individual neurons, giving a vastly more precise picture than anything achievable by other means. In other respects the experiments broadly followed the design of Libet's own, using a similar clock-face approach to measure the time when subjects felt they decided to press a button. Libet discovered that a Readiness Potential (RP) could be detected as much as half a second before the subject was conscious of deciding to move; the new experiments show that data from a population of 250 neurons in the SMA (the Supplementary Motor Area) were sufficient to predict the subject's decision 700 ms in advance of the subject's own awareness, with 80% accuracy.

The more detailed picture which these experiments provide helps clarify some points about the relationship between pre-SMA and SMA proper, and suggest that the sense of decision reported by subjects is actually the point at which a growing decision starts to be converted into action, rather than the beginning of the decision-forming process, which stretches back further. This may help to explain the results from fMRI studies which have found the precursors of a decision much earlier than 500 ms beforehand. There are also indications that a lot of the activity in these areas might be more concerned with suppressing possible actions than initiating them - a finding which harmonises nicely with Libet's own idea of 'free won't' - that we might not be able to control the formation of impulses to act, but could still suppress them when we wanted.

For us, though, the main point of the experiments is that they appear to provide a strong vindication of Libet and make it clear that we have to engage with his finding that our decisions are made well before we think we're making them.

What are we to make of it all then? I'm inclined to think that the easiest and most acceptable way of interpreting the results is to note that making a decision and being aware of having made a decision are two different things (and being able to report the fact may be yet a third). On this view we first make up our minds; then the process of becoming aware of having done so naturally takes some neural processing of its own, and hence arrives a few hundred milliseconds later.

That would be fine, except that we strongly feel that our decisions flow from the conscious process, that the feelings we are aware of, and could articulate aloud if we chose, are actually decisive. Suppose I am deciding which house to buy: house A involves a longer commute while house B is in a less attractive area. Surely I would go through something like an internal argument or assessment, totting up the pros and cons, and it is this forensic process in internal consciousness which causally determines what I do? Otherwise why do I spend any time thinking about it at all - surely it's the internal discussion that takes time?

Well, there is another way to read the process: perhaps I hold the two possibilities in mind in turn: perhaps I imagine myself on the long daily journey or staring at the unlovely factory wall.

Which makes me feel worse? Eventually I get a sense of where I would be happiest, perhaps with a feeling of settling one alternative and so of what I intend to do. On this view the explicitly conscious part of my mind is merely displaying options and waiting for some other, feeling part to send back its implicit message. The talky, explicit part of consciousness isn't really making the decision at all, though it (or should I say 'I'?) takes responsibility for it and is happy to offer explanations.

Perhaps there are both processes in involved in different decisions to different degrees. Some purely rational decisions may indeed happen in the explicit part of the mind, but in others - and Libet's examples would be in this category - things have to feel right. The talky part of me may choose to hold up particular options and may try to nudge things one way or another, but it waits for the silent part to plump.

Is that plausible? I'm not sure. The willingness of the talky part to take responsibility for actions it didn't decide on and even to confect and confabulate spurious rationales, is very well established (albeit typically in cases with brain lesions), but introspectively I don't like the idea of two agents being at work I'd prefer it to be one agent using two approaches or two sets of tools - but I'm not sure that does the job of accounting for the delay which was the problem in the first place...

Lies, Damned Lies and Statistics

5 June 2011

There has been quite a furore over some remarks made by Chomsky at the recent MIT symposium Brains, Minds and Machines; Chomsky apparently criticised researchers in machine learning who use purely statistical methods to produce simulations without addressing the meaning of the behaviour being simulated - there's a brief account of the discussion in Technology Review. The wider debate was launched when Google's Director of Research, Peter Norvig issued a response to Chomsky. So far as I can tell the point being made by Chomsky was that applications like Google Translate, which uses a huge corpus of parallel texts to produce equivalents in different languages, do not tell us anything about the way human beings use language.

My first reaction was surprise that anyone would disagree with that. The weakness of translation programmes is that they don't deal with meanings, whereas human language is all about meanings. Google is more successful than earlier attempts mainly because it uses a much larger database. We always knew this was possible in principle, and that there is no theoretical upper limit to how good these translations can get (in theory we can have a database of everything ever written); it's just that we (or I, anyway) underestimated how much would be practical and how soon.

As an analogy, we could say it's like collecting exhaustive data about the movement of heavenly bodies. With big enough tables we could make excellent predictions about eclipses and other developments without it ever dawning on us that the whole thing was in fact about gravity. Norvig reads into Chomsky's remarks a reassertion of views expressed earlier that 'statistical' research methods are no more than 'butterfly collecting'. It would be a little odd to dismiss the collection of data as not really proper science - but it's true that we celebrate Newton who set out the law of gravity, and not poor Flamsteed and the other astronomers who supplied the data.

But hang on there. A huge corpus of parallel texts doesn't tell us anything about the way human beings use language? That can't be right, can it? Surely it tells us a lot about the way certain words are used, and their interpretation in other languages? Norvig makes the point that all interpretation is probabilistic - we pick out what people most likely meant. So probabilistic models may well enlighten us about how human beings process language. He goes on to point out that there may be several different ways of capturing essentially the same theory, some more useful for some purposes than others: why rule out statistical descriptions?

Hm, I don't know. Certainly there are times when we have to choose between rival interpretations, but does that make interpretation essentially probabilistic? There are times when we have to choose between two alternative jigsaw pieces, but I wouldn't say that solving a jigsaw was a statistical matter. Even if we concede that human interpretation of language is probabilistic that glosses over the substantial point that in human beings the probabilistic judgements are about likely meanings, not about likely equivalent sentences. Human language is animated by intentionality, by meaning things: and unfortunately at the moment we have little idea of how intentionality works.

But then, if we don't know how meaning works, isn't it possible that it might turn out to be statistical (or probabilistic, or anyway some sort of numerical)? I don't think it is possible to rule this out. If we had some sort of neural network which was able to do translations, or perhaps talk to us intelligently, we might be tempted to conclude that its inner weightings effectively encoded a set of probabilities about sentences and/or a set of meanings - mightn't we? And who knows, by examining those weightings might we not finally glean some useful insight into how leaden numbers get transmuted into the gold of semantics?

As I say, I don't think it's possible to rule that out - but we know how Google Translate works and it isn't really anything like that. Norvig wouldn't claim -would he? - that Google Translate or other programs written on a similar basis display any actual understanding of the sentences they process. So it still seems to me that, whether or not his wider attitudes are well-founded, Chomsky was essentially right.

To me the discussion has several curious echoes of the Chinese Room. I wonder whether Norvig would say that the man in the room understands Chinese?

Most Human

25 June 2011

In The Most Human Human: A Defence of Humanity in the Age of the Computer, Brian Christian gives an entertaining and generally sensible account of his campaign to win the 'most human human' award at the Loebner prize.

As you probably know, the Loebner prize is an annual staging of the Turing test: judges conduct online conversations with real humans and with chatbots and try to determine which is which. The main point of the contest is for the chatbot creators to deceive as many judges as they can, but in order to encourage the human subjects there's also an award for the participant considered to be least like a computer. Christian set himself to win this prize, and intersperses the story of his effort with reflections on the background and significance of the enterprise.

He tries to ramp up the drama a bit by pointing out that in 2008 a chatbot came close to succeeding, persuading 3 of the judges that it was human: four would have got it the first-ever win. This year, therefore, he suggests, might in some sense prove to be humanity's last stand. I think it's true that Loebner entrants have improved over the years. In the early days none of the chatbots was at all convincing and few of their efforts at conversation rose above the ludicrous – in fact, many early transcripts are worth reading for the surreal humour they inadvertently generated. Nowadays, if they're not put under particular pressure the leading chatbots can produce a lengthy stream of fairly reasonable responses, mixed with occasional touches of genius and – still – the inevitable periodic lapses into the inapposite or plain incoherent. But I don't think the Turing Test is seriously about to fall. The variation in success at the Loebner prize has something to do with the quality of the bots, but more with the variability of the judges. I don't think it's made very clear to the judges what their task is, and they seem to divide into hawks and doves: some appear to feel they should be sporting and play along with the bots, while others approach the conversations inquisitorially and do their best to catch their 'opponents' out. The former approach sometimes lets the bots look good; the latter, I'm afraid, never really fails to unmask them. I suspect that in 2008 there just happened to be a lot of judges who were ready to give the bots a fighting chance.

How do you demonstrate your humanity? In the past some human contenders have tried to signal their authenticity with displays of emotional affect, but I should have thought that if that approach was susceptible to fakery. However, compared to any human being, the bots have little information and no understanding. They can therefore be thrown off either by allusions to matters of fact that a human participant would certainly know (the details of the hotel breakfast; a topical story from the same day's news) but would not be in the bots' databases; or they can be stumped by questions that require genuine comprehension (prhps w cld sk qstns wth n vwls nd sk fr rpls n the sm frmt?). In one way or another they rely on scripts, so as Christian deduces, it is basically a matter of breaking the pattern.

I'm not altogether sure how it comes about that humans can break pattern so easily while remaining confident that another human will readily catch their drift. Sometimes it's a matter of human thought operating on more than one level, so that where two topics intersect we can

leap from one to the other (a feature it might be worth trying to build into bots to some extent). In the case of the Loebner, though, hawkish judges are likely to make a point of leaving no thread of relevance whatever between each input, confident that any human being will be able to pick up a new narrative instantly from a standing start. I think it has something to do with the un-named human faculty that allows us to deal with pragmatics in language, evade the frame problem, and effortlessly catch and attribute meanings (at least, I think all those things rely at least partly on a common underlying faculty,or perhaps on an unidentified common property of mental processes).

Christian quotes an example of a bot which appeared to be particularly impoverished, having only one script: if it could persuade the judge to talk about Bill Clinton it looked very good, but s soon as the subject was changed it was dead meat. The best bots, like Rollo Carpenter's Jabberwacky, seem to have a very large repertoire of examples of real human responses to the kind of thing real humans say in a conversation with chatbots (helpfully real humans are not generally all that original in these circumstances, so it's almost possible to treat chatbot conversation as a large but limited domain in itself). They often seem to make sense, but still fall down on consistency, being liable to give random and conflicting answers, for example about their own supposed gender and marital status.

Reflecting on this, Christian notes that a great deal of ordinary everyday human interaction effectively follows scripts, too. In the routine of small talk, shopping, or ordering food, there tend to be ritual formulas to be followed and only a tiny exchange of real information. Where communication is poor, for example where there is no shared language, it's still often relatively easy to get the tiny nuggets of actual information across and complete the transaction successfully (though not always: Christian tells against himself the story of a small disaster he suffered in Paris through assuming, by analogy with Spanish, that 'station est' must mean 'this station').

Doesn't this show, Christian asks, that half the time we're wasting time? Wouldn't it be better if we dropped the stereotyped phatic exchanges and cut to the chase? In speed-dating it is apparently necessary to have rules forbidding participants to ask certain standard questions (Where do you live? What do you do for a living?) which eat up scarce time without people getting any real feel for each other's personality. Wouldn't it be more rewarding if we applied similar rules to all our conversations?

This, Christian thinks, might be the gift which artificial intelligence ultimately bestows on us. Unlike some others, he's not worried that dialogue with computers will make us start to think of ourselves as machines – the difference, he thinks, is too obvious. On the contrary, the experience of dealing with robots will bring home to us for the first time how much of our own behaviour is needlessly stereotyped and robotic and inspire us to become more original – more human – than we ever were before.

In some ways this makes sense. As a similar point it has sometimes occurred to me in the past to wonder whether our time is best spent by so many of us watching the same films and reading the same books. Too often I have had conversations sharing identical memories of the same television programme and quoting the same passages from reviews in the same newspapers. Mightn't it be more productive, mightn't we cover more ground, if we all had different experiences of different things?

Maybe, but it would be hard work. If Christian had his way, we should no longer be saying things like this.¶

- Hi, how are you?
- Fine, thanks, you too?
- Yeah, not so bad. I see the rain's stopped.
- Mm, hope it stays fine for the weekend.
- Oh, yeah, the weekend.
- Well, it's Friday tomorrow - at last!
- That's right. One more day to go.

Instead I suppose our conversations would be earnest, informative, intense, and personal.¶

- Tell me something important.
- Sometimes I'm secretly glad my father died young rather than living to see his hopes crushed.
- Mithridates, he died old.
- Ugh: Housman's is the only poetry which is necessarily improved by parody.
- I've never respected your taste or your intellect, but I've still always felt protective towards you.
- There's a useful sociological framework theory of amorance I can summarise if it would help?
- What are you really thinking?

Perhaps the second kind of exchange is more interesting than the first, but all day every day it would be tough to sustain and wearing to endure. It seems to me there's something peculiarly Western about the idea that even our small talk should be made to yield a profit. I believe historically most civilisations have been inclined to believe that the world was gradually deteriorating from a previous Golden Age, and that keeping things the way they had been in past was the most anyone could generally aspire to. Since the Renaissance, perhaps, we have become more accustomed to the idea of improvement and tend to look restlessly for progress: a culture constantly gearing up and apparently preparing itself for some colossal future undertaking the nature of which remains obscure. This driven quality clearly yields its benefits in prosperity for us, but when it gets down to the personal level it has its dangers, at worst it may promote slave-like levels of work, degrade friendship into networking and reinterpret leisure as mere recuperation. I'm not sure I want to see self-help books about leveraging those moments of idle chat. (In fairness, that's not what Christian has in mind either.)

Christian may be right, in any case, that human interaction with machines will tend to emphasise the differences more than the similarities. I won't reveal whether he ultimately succeeded in his quest to be Most Human Human (or perhaps was pipped at the post when a rival and his judge stumbled on a common and all-too-human sporting interest?), but I can tell you that this was not on any view humanity's last stand: the bots were routed.

A Unified Theory of Consciousness

10 July 2011

This paper on 'Biology of Consciousness' embodies a remarkable alliance: authored by Gerald Edelman, Joseph Gally, and Bernard Baars, it brings together Edelman's Neural Darwinism and Baars' Global Workspace into a single united framework. In this field we're used to the idea that for every two authors there are three theories, so when a union occurs between two highly-respected theories there must be something interesting going on.

As the title suggests, the paper aims to take a biologically-based view, and one that deals with primary consciousness. In human beings the presence of language among other factors adds further layers of complexity to consciousness; here we're dealing with the more basic form which, it is implied, other vertebrates can reasonably be assumed to share at least in some degree. Research suggests that consciousness of this kind is present when certain kinds of connection between thalamus and cortex are active: other parts of the brain can be excised without eradicating consciousness. In fact, we can take slices out of the cortex and thalamus without banishing the phenomenon either: the really crucial part of the brain appears to be the thalamic intralaminar nuclei. Why them in particular? Their axons radiate out to all areas of the cortex, so it seems highly likely that the crucial element is indeed the connections between thalamus and cortex.

The proposal in a nutshell is that groups of neurons dispersed but re-entrantly connected in this way constitute a flexible Global Workspace where different inputs can be brought together, and that this is the physical basis of consciousness. Given the extreme diversity and variation of the inputs, the process cannot be effectively ringmastered by a central control, but instead the contents and interactions are determined by a selective process - Edelman's neural Darwinism (or neural group selection): developmental selection ('fire together, wire together'), experiential selection, and co-ordination through re-entry.

This all seems to stack up very well (almost too sensible to be the explanation for anything as strange as consciousness). The authors note that it helps explain the unity of consciousness. It might seem that it would be useful for a vertebrate to be able to pay attention to several different inputs at once, but it doesn't seem to work that way, and perhaps it's because there is only one 'Dynamic Core'. This constraint must have compensating advantages, and the authors suggest that these may lie in the ability of a single piece of data to be reflected quickly across a whole raft of different sub-systems. I don't know whether that is an explanation, but I suspect the real reason is to do with outputs rather than inputs. It might seem useful to deal with more than one input at a time, but having more than one plan of action is response has obvious negative survival value. It seems entirely plausible that part of the value of a Global Workspace would come from its role in filtering down multiple stimuli towards a single coherent set of actions. The authors reckon, in fact, that linked changes in the core could give rise to a coherent flow of discriminations which could account for the 'stream of consciousness'. I'm not altogether sure about that - without saying it's impossible a selective process without central control can give rise to the kind of intelligible flow we experience in our mental processes, I don't quite see how the trick is done. Darwin's original brand of evolution, after all, gave rise to speciation, not coherence of development. But no doubt much more could be said about this.

Thus far, we seem on pretty solid ground. The authors note that they haven't accounted for certain key features of consciousness, in particular subjective experience and the sense of self: they also mention intentionality, or meaningfulness. These are, as they say, non-trivial matters and I think honour would have been satisfied if the paper concluded there: instead however, the authors gird their loins and give us a quick view of how these problems might in their view be vanquished.

They start out by emphasising the importance of embodiment and the context of the 'behavioural trinity' of brain, body, and world. By integrating sensory and motor signal with stored memories, the 'Dynamic Core' can, they suggest, generate conceptual content and provide the basis for intentionality. This might be on the right track, but it doesn't really tell us what concepts are or how intentionality works: it's really only an indication of the kind of theory of intentionality which, in a full account, might occupy this space.

On subjective experience, or qualia, the authors point out that neural and bodily responses are by their nature private, and that no third-person description is powerful enough to convey the actual experience. They go on to deny that consciousness is causal: it is, they say, the underlying neural events that have causal power. This seems like a clear endorsement of epiphenomenalism, but I'm not clear how radical they mean to be. One interpretation is that they're saying consciousness is like the billows: what makes the billows smooth and bright? Well, billows may be things we want to talk about when looking at the surface of the sea, but really if we want to understand them there's no theory of billows independent of the underlying hydrodynamics. Billows in themselves have no particular explanatory power. On the other hand, we might be talking about the Hepplewhiteness of a table. This particular table may be Hepplewhite, or it may be fake. Its Hepplewhiteness does not affect its ability to hold up cups; all that kind of thing is down to its physical properties. But at a higher level of interpretation Hepplewhiteness may be the thing that caused you to buy it for a decent sum of money. I'm not clear where on this spectrum the authors are placing consciousness.

On the self, the authors suggest that neural signals about one's own responses and proprioception generate a sense of oneself as a separate entity: but they do not address the question of whether and in what sense we can be said to possess real agency: the tenor of the discussion seems sceptical, but doesn't really go into great depth. This is a little surprising, because the Global Workspace offers a natural locus in which to repose the self. It would be easy, for example, to develop a compatibilist theory of free will in which free acts were defined as those which stem from processes in the workspace but that option is not explored.

The paper concludes with a call to arms: if all this is right, then the best way to vindicate it might be to develop a conscious artefact: a machine built on this model which displays signs of consciousness - a benchmark might be clear signs of the ability to rotate an image or hold a simulation.'

The authors acknowledge that there might be technical constraints, but I think they an afford to be optimistic. I believe Henry Markram, of the Blue Brain project, is now pressing for the construction of a supercomputer able to simulate an entire brain in full detail, so the construction of a mere Global Dynamic Core Workspace ought to be within the bounds of possibility, if there are any takers?

Thatter Way To Consciousness

24 July 2011

Aping Mankind' is a large scale attack by Raymond Tallis on two reductive dogmas which he characterises as 'Neuromania' and 'Darwinitis'. He wishes especially to refute the identification of mind and brain, and as an expert on the neurology of old age, his view of the scientific evidence carries a good deal of weight. It's a vigorous, useful, and readable contribution to the debate.

Tallis persuasively denounces exaggerated claims made on the basis of brain scans, notably claims to have detected the 'seat of wisdom' in the brain. These experiments, it seems, rely on what are essentially fuzzy and ambiguous pictures arrived at by subtraction in very simple experimental conditions, to provide the basis for claims of a profound and detailed understanding far beyond what they could possibly support. Of course, the fact that some claims to have reduced thought to neuronal activity are wrong does not mean that thought cannot nevertheless turn out to be neuronal activity, but Tallis pushes his scepticism a long way. Most people, I think, would accept that Wilder Penfield's classic experiments, in which the stimulation of parts of the brain with an electrode caused an experience of remembered music in the subject, pretty much show that memories are encoded in the brain; but Tallis does not accept that neurons could constitute memories. For memory you need a history, you need to have formed the memories in the first place, he says: Penfield's electrode was merely reactivating memories which already existed.

Tallis seems to be basing his view on a kind of Brentanoesque incredulity about the utter incompatibility of the physical and the mental. Some of his arguments have a refreshingly simple (or if you prefer, naive) quality: when we experience yellow, he points out, our nerve impulses are not yellow. True enough, but then a word need not be printed in yellow ink to encode yellowness either. Tallis quotes Searle offering a dual-aspect explanation: water is H2O, but H2O molecules do not themselves have watery properties: in the same way our thoughts can be neural activity without the neurons themselves resembling thoughts. Tallis utterly rejects this: he maintains that to have different aspects requires a conscious observer, so we're smuggling in the very thing we need to explain. I think this is an odd line to take. If things don't have different aspects until an observer is present, what determines the aspects they eventually have? If it's the observer, we seem to slipping towards idealism or solipsism, which I'm sure Tallis would not find congenial; if it's the thing that determines its own aspects, didn't it really sort of have them all along? Tallis points out that things can have an indefinite number of different aspects, but so what? Tallis seems to be adopting the view that an appearance can only properly be explained by another thing that already has that same appearance. It must be clear that if we take this view we're never going to get very far with our explanations.

But I think that's the weakest point in a sceptical case which is otherwise fairly plausible. Tallis is Brentanoesque in another way in that he emphasises the importance of intentionality - quite rightly, I think. In his view the capacity to think explicitly about things is a key unique feature of human mindfulness, and that may well be correct. I'm less sure about his characterisation of intentionality as an outward arrow. Perception, he says, is usually represented purely in terms of

information flowing in, but there is also a corresponding outward flow of intentionality. The rose we're looking at it hits our eye (or rather a beam of light from the rose does so), but we also, as it were, think back at the rose. Is this a useful way of thinking about intentionality? It has the merit of foregrounding it, but I think we'd need a theory of intentionality in order to judge whether talk of an outward arrow was helpful or confusing, and no theory is on offer.

Tallis has a very vivid evocation of a form of the binding problem, the issue of how all our different sensory inputs are brought together in the mind coherently. As normally described, the binding problem seems like lip-synch issues writ large: Tallis focuses instead on the strange fact that consciousness is united and yet composed of many small distinct elements at the same time. He rightly points out that it's no good having a theory which merely explains how things are all brought together: if you combine a lot of nerve impulses into one you just mash them. I think the answer may be that we can experience a complex unity because we are complex unities ourselves, but in any case it's an excellent and thought-provoking exposition.

Tallis' attack on' Darwinitis' takes on Cosmidoobianism, memes and the rest with predictable but nonetheless enjoyable vigour. Again, he presses things quite a long way. It's one thing to doubt whether every feature of human culture is determined by evolution: Tallis seems to suggest that human culture has no survival value, or at any rate, had none until recently, too recently to account for human development. This is rather reminiscent of the argument put by Alfred Russel Wallace, Darwin's co-discoverer, who said that evolution could not account for human intelligence because a caveman could have lived his life perfectly well with a much less generous helping of it. The problem is that this leaves us needing a further explanation of why we are so brainy and cultured; Wallace, alas, ended up resorting to spiritualism to fill the gap. It seems better to me to draw a clear distinction between the capacity for human culture, which is wholly explicable by evolutionary pressure, and the contents of human culture, which are at least partly ephemeral, variable, and non-hereditary.

Tallis points out that some sleight of hand with technology is not unknown in this area: a word implying full mental activity is used metaphorically - a 'smart' bomb is said to be 'hunting down' its target - and the important difference is covertly elided. He notes the particular slipperiness of the word 'information', something we've noted before.

It is a weakness of Tallis' position that he has no general alternative theory to offer in place of those he is attacking - consciousness remains a mystery (he sympathises with Colin McGinn's mysterianism to some degree, incidentally,but reproves him for suggesting that our inability to understand ourselves might be biological). However, he does offer positive views of selfhood and free will, both of which he is concerned to defend. Rather than the brain, he chooses to celebrate the hand as a defining and influential human organ: opposable thumbs allow it to address itself and us: it becomes a proto-tool and this gives us a sense of ourselves as acting on the world in a tool-like way. In this way we develop a sense of ourselves as a distinct entity and an agent, and existential intuition. This is OK as far as it goes though it does sound to some extent like another theory of how we get an impression, or dare we say an illusion, of selfhood and agency, the very thing Tallis wants to refute. In response to critics who have pointed to the elephant's trunk and the squid's tentacles, Tallis grudgingly concedes that hands alone are not all you need.

Turning to free will, Tallis tackles Libet's experiments (which seem to show that a decision to move one's hand is measurably made before one becomes aware of it). So, he says, the decision to move the hand can be tracked back half a second? Well, that's nothing: if you like

you can track it back days, to when the experimental subject decided to volunteer; moreover, the aim of the subject was not just to move the hand, but also to help that nice Dr Libet, or to forward the cause of science. In this longer context of freely made decisions the precise timing of the RP is of no account.

To be free according to Tallis, an act must be expressive of what the agent is, the agent must seem to be the initiator, and the act must deflect the course of events. If we are inclined to doubt that we can truly deflect the course of events, he again appeals to a wider context: look at the world around us, he says, and who can doubt that collectively we have diverted the course of events pretty substantially? I don't think this will convert any determinists. The curious thing is that Tallis seems to be groping for a theory of different levels of description, or dare I say it, a dual aspect theory. (I would have thought dual-aspect theories ought to be quite congenial to Tallis, as they represent a rejection of 'nothing but' reductionism in favour of an attempt to give all levels of interpretation parity of esteem.)

As I say, there is no new theory of consciousness on offer here, but Tallis does review the idea that we might need to revise our basic ideas of how the world is put together to accommodate it. He is emphatically against traditional dualism, and he firmly rejects the idea that quantum physics might have the explanation too. Panpsychism may have a certain logic but generate more problems than it solves. Instead he points again to the importance of intentionality: 'Thatter' may be as important as matter.

Closing the Gap

30 Aug 2011

One way of setting up the vexed question of qualia is to claim that there is an explanatory gap between what science tells us about our sensory organs and nervous system on the one hand, and actual real experience on the other. Nothing in the biological/physical story, it's claimed, tells us what the redness of a rose or the smell of violets is actually like, and nothing of that kind ever could. The two aspects of the experience do not connect with each other. In the latest JCS, Michael Pauen sets out to show us that that supposed gap does not exist.

He attacks from four angles. First, the thought-experiments put forward to point out the gap are inconsistent: second, they require the first-person view to have a special privilege which it hasn't; third, the arguments for the gap are circular, resting on the same intuition they set out to vindicate. Finally, he offers a historical argument to show that apparent gaps of a similar kind have disappeared in the past as our scientific understanding grew: there's no reason to think that this one too will stop seeming plausible when we understand things better.

We've often discussed the thought-experiments in question. There's Mary, brought up in monochrome but knowing all the science, who nevertheless knows something more, it's claimed, when she sees what redness is really like. There are the 'zombies', creatures exactly like us in every physical detail and consequently in behaviour too, yet having no real experiences: the possibility of such beings, if you accept it, proving that there's more than just physics going on. There are also the many variants of the inverted spectrum, where what I experience when I see blue is what you experience when you see red; our inability to discover or communicate this difference once again proving the existence of the gap.

What's inconsistent about all that? Pauen introduces a new kind of zombie (Another new kind of zombie? There must be a dozen kinds in the literature already.) . These are part-time zombies. Some of the time they do have subjective experience just like us, at other times, as it were, the lights go out for a while. The intermittent nature of their experience doesn't affect their behaviour, of course, because ex hypothesi zombies are all exactly like their non-zombie equivalents in all physical matters. Moreover, because their zomboid episodes leave no physical traces in their brains, they can't remember whether or not they had experiences at any given time. For that matter, we can't tell whether we are part-time zombies ourselves, because we wouldn't remember in either case. In fact, we can't tell whether we ever did have subjective experience: so the theory of the gap ultimately undermines our reasons for believing in the gap in the first place.

This is a useful argument: still, I think Pauen may underestimate how far committed dualists might be prepared to go in digging in on this issue. Why has the experience got to leave physical traces in the brain, they might ask - I might just remember it spiritually. Pauen in reply would no doubt point out that such a spiritual memory could never have physical effects, so nothing they say with a physical mouth or write with a physical hand could ever have been caused by those ineffable experiences. This is certainly an uncomfortable position for the dualists to be in, but it really only expands and dramatises the bind they were already in about the acausal nature of

qualia, and you could argue that it's only a little worse than the problems about interaction with the physical world which dualists have always faced. Somehow they seem to live with it.

Pauen's second argument is meant to show that there's no privileged first-person access to qualia: what does that actually mean? It's obvious that you can't, strictly, see things from my perspective without being me: but the explanatory gap also requires that there are facts about my experience (facts, presumably, about what it is like) that you can never get to know from the third-person perspective (if you could get to know everything from the third-person view there would be no gap). Pauen argues that if the subjects can recognise qualia, there must be some resulting difference in their epistemic or functional dispositions: these in turn will be recognisable from the third-person point of view. Contrariwise, if there is no functional difference, the subjects themselves will be unable to recognise the qualia, because the mere ability to do so would itself constitute a functional difference. Although I think he's right about this, I think he has again assumed more agreement than some would be prepared to give. I suppose most people these days are broadly functionalist in a general sense, but not everyone would accept that the ability to recognise the presence or absence of qualia has anything to do with functional dispositions, if that is to be read as referring to functional dispositions of physical matter. It's tough to see how it could be otherwise, but there are mouths ready to bite that bullet.

Why is that? Why do people live with positions that involve such bad, unresolved philosophical problems? I think the dualists might say: look Michael, we know this is severely problematic - they don't call it the hard problem for nothing - but what can we do? These qualia are right there, immediately obvious. It's as if you'd come up with some really cogent arguments to show we had no heads, arguments we couldn't fault. In those circumstances we're forced to say, sorry Michael, but we just have not been decapitated - we just haven't - so for the time being we're going to have to work on the assumption that there's something wrong with your case for acephaly even though just at the moment we can't give you an unproblematic alternative.

Ah, but that plays into Pauen's hands, because his third argument addresses the strength of the intuition behind qualia. He points out, quite rightly, I think, that we generally start with a strong feeling that there is something it is like; then the various arguments get presented to validate this intuition. But on examination all of the arguments rest in the end on the same intuition. It is logically open to us to just deny that Mary learns anything new, to deny the possibility of zombies, and dismiss the issue of the inverted spectrum as meaningless. No argument compels us to do otherwise, it's just that the same basic intuition is supposed to lead us in the other direction. Intuitions, Pauen points out, don't prove anything.

That's right of course, but there are a couple of things to be said. The first is that I don't think the advocates of qualia actually suppose they're putting forward a tight logical case. The arguments are, as it were, ostensive, they don't deduce the existence of qualia, they simply display it. What we're offered is not really arguments so much as what Dennett calls 'intuition pumps', less powerful but legitimate tools if used properly.

Second, in calling them 'intuitions', Pauen sells these impressions a bit short. There are intuitions and intuitions: here we're not talking about a hunch that people have qualia, we're talking about having qualia right in your face, having them in the most direct and unmediated manner - a manner which some might even argue had a special immunity from error (I can be wrong about the fact that I'm seeing a rose, but can't be wrong about the fact that I seem to be seeing a rose (Eric Schwitzgebel might have something to say about that though)). For those

who believe in them, qualia may be ineffable, but they're also undeniable in a unique way denied to mere feelings about things are likely to be.

Or so you may think: but Pauen's last argument is the historical one that things which one seemed irreducible to us have often ended up being perfectly explicable once we knew a bit more about the underlying science. In general Pauen is arguing against the view that there is an explanatory gap in principle, but here his argument also implicitly rebuts those who, like Colin McGinn, never claimed qualia were mysterious in themselves, only forever mysterious to us. I don't suppose anyone ever changed their mind because the opposition's retelling of events convinced them they were on the wrong side of history, but I found Pauen's account, which takes up a substantial part of the paper, enlightening. Besides Fechner and Du Bois-Reymond he picks out Descartes for particular attention, and it was refreshing to see him given fair and accurate treatment for once instead of being blamed for imaginary theatres and what have you.

On this general historical argument I again think Pauen is basically right. It is already possible to explain many aspects of vision in ways which tend to reduce the sense of ineffability a bit, and no doubt this will continue to develop. But that may not do the job of dispelling all the magic. Part of the sensation of mystery about qualia, I suspect, does come from a certain mere bogglement over the difference between first and third person view; and a large part comes from the inexplicable haecceity of the world and even worse, of ourselves. These things are not going to stop bothering people any time soon.

When Incredulities Meet

4 September 2011

Sometimes mistakes can be more interesting than getting it right. Last week I was thinking about Pauen's claim, reasonable enough, that belief in qualia is ultimately based on the intuitive sense that experience and physics are two separate realms. The idea that subjective stuff, the redness of red and so on, could be nothing but certain jigs danced by elementary particles, provokes a special incredulity. What's the famous quote that sums that up, I thought? Something about…

> *This phenomenal quality is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it.*

That captures the incredulity quite nicely. However, it dawned on me that it was Brentano, and he didn't say 'phenomenal quality', he said 'intentional inexistence'.

So it turns out we have two incredulitites, one about qualia - subjectivity or 'what it is like', one about intentionality - 'aboutness' or meaningfulness. To me, they have a very similar feel. So what do we say about that? I can see four reasonable possibilities.

1. The resemblance is superficial: just because your mind boggles at two different things, it doesn't mean the two things are identical.
2. The incredulity is the same because it's not specifically attached to qualia or intentionality, it's just characteristic of mental phenomena of all kinds.
3. The incredulity arises from intentionality, and qualia have it because they are intentional in nature.
4. The incredulity arises from qualia, and intentionality has it because it arises out of qualia.

Although 1. is a very rational line to take, I can't help feeling there is at least a little more to it than that. I don't detect in myself a third incredulity - I don't feel that nothing in subjectivity could possibly account for intentionality, or vice versa: that remains to be examined. And to put it no higher, it would be nice if we could tidy things up by linking the two problems, or even perhaps reducing one to the other. One inexplicable realm is bad enough.

I suppose 2. is what Brentano himself might have said. I don't know whether we'd now be quite so quick to bestow the mystery on all mental phenomena: it doesn't seem so implausible now that calculation or choosing a chess move might be nothing more than a special kind of physical activity. Moreover, if the problem doesn't come from intentionality or qualia, we seem to have a third problem distinct from either, which is unwelcome, and a slight difficulty over the relationships. It doesn't seem much of an answer to say that qualia seem strange and non-physical because they're mental, unless we can go on to say a lot more about the spookiness of the mental and why it attaches to subjective experience the way it does.

I suppose we could go dualist here and say that mental things exist in a separate domain in which both qualia and meanings participate. Isn't something like that the main reason dualists are dualists, in fact? Taking that route involves the usual problems of explaining the interaction

between worlds and indeed, giving some explanation of how the second world works. If we don't give that latter explanation we seem only to have deferred the issues.

It might be easier if we said that the mental is essentially a different level of explanation within an essentially monist universe. For me, that looks at least a starter so far as intentionality goes, but not for qualia. They're not explanatory at all, quite the reverse. This brings out some interesting differences. In the case of qualia we already have a pretty full scientific account of how the senses work. We pretty much know what we'd reduce qualia to, if we're in the market for a reduction. In a sense, the way is clear: there's no work in the ordinary world that we need qualia to do, we just need an extra ineffable zing from somewhere, something we could arguably dispense with. For intentionality, things are much worse. There is no scientific account of meaning, we don't really know how the brain deals with it, yet it is an essential part of our lives which can't be dismissed as airy-fairy obscurantism. Curiously, of course, it's qualia which are seen as the hard Problem, while intentionality is part of the easy one. I suppose this is because when we contemplate intentionality, it doesn't seem intractable. We may not know how it works, but it looks like the kind of thing we could get a grip on given a couple of insights; whereas there seems no way of scaling the smooth glassy wall presented by qualia.

Here's a thought: if we're saying that the two issues are different facets of the same problem, we ought to be able to apply the established qualia arguments to intentionality. We can't do it the other way because I don' t think there are any arguments for the existence of intentionality - nobody denies it.

So: the zombies go quite well, I think: we'd say that intentionality zombies (another kind - sorry) look and behave like us, but never actually mean what they say or understand the words they read. By some process they come out with appropriate responses, but in the same sort of sense as the original zombies, the lights are all out.

Then instead of inverted spectra, we'd have inverted meanings. This is trickier, because there's no tidy realm of meaning equivalent to the spectrum we can use - unless we co-opt the spectrum itself and say that when you mean red, you say blue… That doesn't seem to work. Could we say that you actually mean the negation of everything you say, but for some reason act otherwise…? Maybe not.

All right, let's try Mary: Intentionality Mary was brought up without ever grasping the meaning of anything, but she understands everything there is to know about cognition… That doesn't seem to make sense.

The problem is always that qualia have no causal effects, whereas meanings and intentions absolutely do: in fact if anything the problem is explaining their efficacy. Noting this, we can see that actually even the zombies didn't really work: we can believe in people who behave like us without having real experience, but it's nonsensical to say that people without desires or intentions would behave the same way unless we're really only talking about some kind of quale of desire or intention.

So if qualia and intentionality are radically different in some respects: these differences might provide at least a hint that 'both mental' is not a good enough explanation for the two incredulities.

What about option 3? Could it be that the incredulity we're concerned with is basically attached to intentionality, and qualia only have it because they are intentional in nature? On the face of it

it seems quite reasonable to think that the redness we experience is about the rose, and that it's the special magic aboutness that adds the extra ineffable quality. With other qualia, though, it's not so clear. If you take happiness to be qualic, what is it about? We can of course be happy about particular things, but that's distinct from just being happy. Moreover, there's plenty of intentionality without qualia: an account book is suffused with intentionality. In fairness, that's only the derived kind - accounts only mean what we make them mean - perhaps it's only the original intentionality of our thoughts that bestows qualicity? But with intentionality, we expect content. We believe and desire and think that x or y, with x or y being capable of expression in words: but it's the whole point of qualia that there's nothing like that available.

Option 4 says qualia are fundamental and intentionality springs from them. John Searle has actually put this view forward (in addition to his view that intentionality is the business of of imposing directions of fit on directions of fit). The suggestion here is that, for example, the feeling of hunger is about food in some basic, primitive sense, and that it's on similar qualia that all our meaningfulness is built. The example has a definite appeal, and there's something attractive about rooting intentionality in the 'three Fs' of survival: making it not some celestial mystery but a particular slant that arises out our nature as competitive and social biological creatures. But there are problems. We must remember that the quale of hunger has no causal effects: it's only the functional counterpart that actually causes us to speak or seek food, so the connection between quale and the expression of beliefs or desires is broken. We may suspect for other reasons that it's not really the quale at work here: the sense in which hunger means food looks very like H.P.Grice's natural meanings (those spots mean measles). Completely inanimate and non-qualic things can have this kind of meaning (those clouds mean rain), so although it is an excellent place to start looking for an analysis of intentionality, it doesn't seem to be a matter of qualia.

Personally, I would reaffirm the view I've often set out before: I haven't a clue what's going on.',

Not Exactly Free

18 September 2011

I see that a piece on nature.com has drawn quite a bit of attention. It provides a round-up of views on the question of whether free will can survive in a post-Libet world, though it highlights more recent findings along similar lines by John-Dylan Haynes and others. The piece seems to be prompted in part by Big Questions in Free Will a project funded by the John Templeton Foundation, which is probably best known for the Templeton Prize, a very large amount of cash which gets given to respectable scientists who are willing to say that the universe has a spiritual dimension, or at any rate that materialism is not enough. BQFW itself is offering funding for theology as well as science: "science of free will ($2.8 million); theoretical underpinnings of free will, round 1 ($165,000); and theology of free will, round 1 ($132,000)". I suppose 'theoretical underpinnings', if it's not science and not theology, must be philosophy; perhaps they called it that because they want some philosophy done but would prefer it not to be done by a philosopher. In certain lights that would be understandable. The presence of theology in the research programme may not be to everyone's taste, although what strikes me most is that it seems to have got the raw end of the deal in funding terms. I suppose the scientists need lots of expensive kit, but on this showing it seems the theologians don't even get such comfortable armchairs as the theorists, which is rough luck.

We have of course discussed the Haynes results and Libet, and other related pieces of research many times in the past. I couldn't help wondering whether, having all this background, I could come up with something on the subject that might appeal to the Templeton Foundation and perhaps secure me a modest emolument? Unfortunately most of the lines one could take are pretty well-trodden already, so it's difficult to come up with an appealing new presentation, let alone a new argument. I'm not sure I have anything new to say. So I've invited a couple of colleagues to see what they can do.


Free will is nonsense; I'm not helping you come up with further 'compatibilist' fudging if that's what you're after. What I can offer you is this: it's not just that Libertarians have the wrong answer, the question doesn't even make sense. The way the *naturenews* piece sets up the discussion is to ask: how can you have free will if the decision was made before you were even aware of it? The question I'm asking is: what the hell is 'you'?

Both Libet's original and the later experiments are cleverly designed to allow subjects to report the moment at which they became aware of the decision: but 'they' are thereby implicitly defined as whatever it is that is doing the reporting. We assume without question that the reporting thing is the person, and then we're alarmed by the fact that some other entity made the decision first. But we could equally well take the view that the silent deciding entity is the person and be unsurprised that a different entity reports it later.

You will say in your typically hand-waving style, I expect, that that can't be right because introspection or your ineffable sense of self or something tells you otherwise. You just feel like you are the thing that does the reporting. Otherwise when words come out of your mouth it wouldn't be you talking, and gosh, that can't be right, can it?

Well, let me ask you this. Suppose you were the decision-making entity, how would it seem to you? I submit it wouldn't seem any way, because as that entity you don't do seeming-to: you just do decisions. You only seem to yourself to have made the decision when it gets seemed back to you by a seeming entity - in fact, by that same reporting entity. In short, because all reports of your mental activity come via the reporting entity, you mistake it for the source of all your mental activity. In fact all sorts of mental processes are going on all over and the impression of a unified consistent centre is a delusion. At this level, there is no fixed 'you' to have or lack free will. Libet's experiments merely tell us something interesting but quite unworrying about the relationship of two current mental modules.

So libertarians ask: do we have free will? I reply that they have to show me the 'we' that they're talking about before they even get to ask that question - and they can't.


Not much of a challenge to come up with something more appealing than that! I've got an idea the Templeton people might like, I think: Dennettian theology.

You know, of course, Dennett's idea of stances. When we're looking to understand something we can take various views. If we take the physical stance, we just look at the thing's physical properties and characteristics. Sometimes it pays to move on to the design stance: then we ask ourselves, what is this for, how does it work? This stance is productive when considering artefacts and living things, in the main. Then in some cases it's useful to move on to the intentional stance, where we treat the thing under consideration as if it had plans and intentions and work out its likely behaviour on that basis. Obviously people and some animals are suitable for this, but we also tend to apply the same approach to various machines and natural phenomena, and that's OK so long as we keep a grip.

But those three stances are clearly an incomplete set. We could also take up the Stance of Destiny: when we do that we look at things and ask ourselves: was this always going to happen? Is this inevitable in some cosmic sense? Was that always meant to be like that? I think you'll agree that this stance sometimes has a certain predictive power: I knew that was going to happen, you say: it was, as it were, SoD's Law.

Now this principle gives us by extrapolation an undeniable God - the God who is the intending, the destining entity. Does this God really exist? Well, we can take our cue from Dennett: like the core of our personhood in his eyes, it doesn't exist as a simple physical thing you can lay your hands on: but it's a useful predictive tool and you'd be a fool to overlook it, so in a sense it's real enough: it's a kind of explanatory centre of gravity, a way of summarising the impact of millions of separate events.

So what about free will? Well of course, one thing you can say about a free decision is that it wasn't destined. How does that come about? I suggest that the RPs Libet measured are a sign of de-destination, they are, as it were, the autopilot being switched off for a moment. Libet himself demonstrated that the impending action could be vetoed after the RP, after all. Most of the time we run on destined automatic, but we have a choice. The human brain, in short, has a unique mechanism which, by means we don't fully understand, can take charge of destiny.

I think my destiny is to hang on to the day job for the time being.

What Are We Even Talking About?

3 October 2011

What do we even mean when we speak of consciousness? As we've noted before, there are many competing and overlapping definitions, and in addition it's pretty clear that the phenomenon itself is complex and that the word refers, in different contexts, to a number of different things.

A few years back, Thomas Natsoulas had a determined go at clarifying the position in his paper "Concepts of Consciousness". For his framework he fell back on the old debating society standby of consulting the dictionary. You might ask whether this was necessarily the best way to go: lexicographers have their own priorities, after all. They typically aim to report the way a word is used; if it's used in ways that are inconsistent or taxonomically incomplete, that isn't a problem for them. On the other hand, the use of dictionary definitions does bring in an element of neutrality, and protects Natsoulas against any charge of skewing his definitions to support his own theoretical views: and the dictionary in question was no less a tome than the complete Oxford English Dictionary (OED), a mighty work of scholarship whose views on almost any subject are not to be lightly dismissed. Natsoulas himself says he sees merit in looking at ordinary, common-sense ideas, which can help remedy the potentially problematic lack of work by psychologists on the conceptual side.

The OED gives six senses of 'consciousness'. The first, which we can call c1, strikes a modern reader as odd: it is *knowing something together*, joint or shared knowledge: *con-scire* as the derivation of the word suggests. There is a definite suggestion of the shared knowledge being a guilty secret, perhaps even an echo of *con-spire.* Although c1 is the ancient sense of the Latin root and seems to have enjoyed a brief revival in the seventeenth century it is no longer current and does not at first seem to offer us much enlightenment on the modern concept. Natsoulas, however, points out that it captures the idea of consciousness as a social, interpersonal thing. He quotes Barlow:

> ...*consciousness is something to do with a relation between brains rather than a property of a single brain*.

Barlow, it seems, went on to suggest that internal consciousness was a kind of 'rehearsal for recounting', which is interesting.

If we doubt that c1 is really important, I suppose we might ask ourselves what the state of mind would be of a human being who never at any stage since birth met another communicative entity. I don't think I'd be ready to say that such a person could not be conscious, but their consciousness would surely be lacking in some important respects.

C2 follows on in a way: it is in effect knowledge shared with oneself, knowing that you know. This sounds like the HOT (Higher Order Theory) and HOP (Higher Order Process) theories which approximately say that a thought is conscious when accompanied by an awareness of that thought. It's also reminiscent of those, like Dennett, who see the internalisation of talking to oneself as the origin of consciousness.

Natsoulas quotes Vygotsky:

> *A function which initially was shared by two people and bore a character of communication between them gradually crystallised and became a means of organisation of the mental life of man himself*

It almost begins to look as if the OED has a rather cogent theory of consciousness.

C3 is awareness, of or that, anything, whether objects in the world or one's own thoughts. Natsoulas insists there must be an object for this form of consciousness: even the thought that 'I am having no thoughts' actually has a content, he points out. Being conscious *without* content is in his eyes properly reserved for c6. He notes that the OED seems to include with c3 a veridicality requirement the claim is that if a man is aware of a bush, but thinks it is a rabbit, he is not really aware of the bush. Natsoulas, rightly I think, disagrees, insisting that even false awareness is still awareness. I think we must certainly preserve the possibility of being aware of something without having to have correct knowledge of its real nature.

I'm not sure whether the indirect awareness of memory falls into this category or the next, but there seems to be an overlap because c4 is, to adopt the OED's Locke quotation:

> *…the perception of what passes in a man's own mind…*

which appears to be a subset of c3. Interestingly the OED seems to think that the primary use of c3 is awareness of one's own mental contents, while using it to mean awareness of actual objects in the world is 'poetic'. Natsoulas concludes that in this respect English has moved on a bit since the OED last looked, but you have to admire the magisterial coherence of the OED view in which consciousness begins by being shared knowledge, becomes knowledge you share with yourself, in which form it is naturally about your own internal states, but by metaphorical extension can also mean your awareness of the external world.

I say c4 seems to be a subset of c3: perhaps as a result, Natsoulas says experimenters are often bedevilled by confusion between the two, claiming that a subject's inability to report a stimulus shows they were never aware of it (whereas the subject can be aware of the stimulus withoug being aware that they are aware).

There might seem to be some dangers in the self-reference of c4, but Natsoulas points out that there's no problem in well-managed higher orders. If there were a ban on higher orders, he argues, introspection could never get properly started.

c5 is not a form of consciousness but rather a set of all the occurrent and previous mental states which putatively make up the individual's existence. This is the sense in which science fiction stories speak of your consciousness being transferred to another body, or to a machine. The OED gives us a quote from Locke:

> *If the same consciousness can be transferr'd from one thinking substance to another, it will be possible that two thinking substances may make but one person…*

Natsoulas is quite happy with the idea that consciousness is not to be identified with the substrate, Locke's 'thinking substance', but he raises some difficulties. What set of states is adequate to constitute a specific consciousness? Must it be all of them? That seems too strong, because if I were to lose one of my memories, I should not thereby lose my identity – I forget things all the time. Perhaps it has to everything I *could* recall – and some forgotten or

unconscious things might still be shaping my mind. Worse, I can remember experiences yet fail to have the feeling that they are really *my* experiences.

Some would regard this defining set of mental states as the ego, or the 'I', which is a way of looking at it: but attempts to use a static central 'I' as the thing that pulls it all together are doomed in Natsoulas' eyes. He thinks that with appropriate care we can use c5 and take account of the vivid and persuasive sense of an inner source without having to grant its ontological reality.

c6 is approximately equivalent to 'awake' and the opposite of unconscious. Searle, in typical commonsensical style, once used c6 as his definition of consciousness:

> *'Consciousness' refers to those states of sentience and awareness that typically begin when we awake from a dreamless sleep and continue until we go to sleep again, or fall into a coma or die or otherwise become 'unconscious'.*

Natsoulas takes this to be the kind of consciousness that has no requirement as to content: you could be conscious in this sense while thinking anything or nothing. In fact although the medical salience of the concept is clear, it seems too open-ended to be of much analytical use. We should certainly be willing to speak of a dog being conscious (or unconscious) in this sense, and I think we'd be willing to push that usage much further – certainly to fish and quite possibly to an ant in certain circumstances. We're not, then, speaking of anything narrowly defined: c6 means something like 'the state when whatever mental activity normally goes on during periods of activity, is going on'.

I think it is open to debate whether this OED-based six-way definition gives us sufficient tools to tackle the probem of consciousness. It does not seem to capture Ned Block's distinction between a- and p-consciousness, more or less the distinction between the targets of the Hard and Easy problem: yet that is one of the most-quoted and used of definitions. I think we might also look for sharper and more useful distinctions between internal and external awareness.

Still, it is a useful exercise, and Natsoulas proceeds to do something with the results, positioning the six senses along four different axes: intersubjectivity, objectivation, apprehension and introspection (this seems to cry out for a diagram, though I appreciate that rendering a four-dimensional space graphically intelligible is a non-trivial matter.

Natsoulas makes little of this concluding exercise, presenting it as a kind of run-through to help get things clear. But it seems obvious that what he's offering is a potential reduction, abstracting away from the six OED definitions to define a consciousness space of four dimensions. Not the least interesting aspect of this is that it implies the conceptual possibility of unknown forms of consciousness which would be situated in unpopulated regions of the space. Suppose, for example, we have a form of consciousness with high intersubjectivity, but low objectivation, apprehension, and introspection? My imagination begins to fail me, but I think that would be a kind of diffuse but powerful general empathy. I'm surprised this aspect has not been explored.

Pain Without Suffering

16 October 2011

The latest issue of the JCS is all about pain. Pain has always been tough to deal with: it's subjective, not a thing out there in the world, and yet even the most hardline reductionist materialist can't really dismiss it as an airy-fairy poetic delusion. We are all intensely concerned about pain, and the avoidance of it is among our most important moral and political projects. When you step back a bit, that seems remarkable: it's easy to see more or less objective reasons why we should want to prevent disease, mitigate the effects of natural disasters, prevent wars and famines - harder to see why near or even at the top of the list of things we care about should be avoiding the occurrence of a particular kind of pattern of neuronal firing.

It's hard even to say what it is. It seems to be a sensation, but a sensation of what? Of.... pain? Our other sensations give us information, about light, sound, temperature, and so on. Pain is often accompanied by feelings of pressure or heat or whatever, but it is quite distinct and separable from those impressions. In itself, the only thing pain tells us is: 'you're in pain'. It seems sensible, therefore, to regard it as not a sensation in the same way as other sensations, but as being something like a kind of deferrable reflex: instead of just moving our arm away from the hot pan it tells us urgently that we ought to do so. So it turns out to be something like a change in our dispositions or a change of weightings in our current projects. That kind of account is appealing except for the single flaw of being evident nonsense. When I'm in the dentist's chair, I'm not feeling a change in my dispositions or anything that abstract, I'm feeling pain - that thing, that bad thing, you know what I mean, even though words fail me.

Is pain actually the most undeniable of qualia? From some angles it looks like it, but qualia are supposed to have no causal effects on our behaviour, and that is exceptionally difficult to believe in the case of pain: if ever anything was directly linked to motivation, pain is it. Undeniability looks more plausible: pain is pre-eminently one of the things it seems we can't be wrong about. I might be mistaken in my belief that my hand has just been sheared off by a saw: that 's a deduction about the state of the world based on the evidence of my senses; I don't see how I could be wrong about the fact that I'm in agony because no reasoning is involved: I just am.

One of the contributors to the JCS might take issue with that, though. S. Benjamin Fink wants to present an approach to difficult issues of phenomenal experience and as his example he offers a treatment of pain which suggests it isn't the simple unanalysable primitive we might think. In Fink's view one of the dangers we need to guard against is the assumption that elements of experience we've always, as it happens, had together are necessarily a single phenomenon. In particular, he wants to argue for the independence of pain and suffering/unpleasantness. Pain, it turns out, is not really bad after all (at least, not necessarily and in itself).

Fink offers several examples where pain and unpleasantness occur separately. An itch is unpleasant but not painful; the burning sensation produced by hot chilis is painful but not

unpleasant (at least, so long as it occurs in the mouths of regular chili eaters, and not in their eyes or a neophyte's mouth). These examples seem vulnerable to a counterargument based on mildness: itches aren't described as pains just because they aren't bad enough; and the same goes for spicy food in a mouth that has become accustomed to it. But Fink's real clincher is the much more dramatic example: pain asymbolia. People with this condition still experience pain but don't mind it. It's not at all that they're anaesthetised: they are aware of pain and can use it rationally to decide when some part of their body is in danger of damage, but they do so , as it were coldly, and don't mind needles being stuck in them for experimental purposes at all. Fink quotes a woman who underwent a lobotomy to cure continual pain: many years later she reported happily that the pain was still there: "In fact, it's still agonising. But I don't mind."

These people are clearly exceptional, but it's worth noting that even in normal people the link between nociception, the triggering of pain-sensing nerve-endings, and the actual experience of pain is by no means as invariable and straightforward as philosophers used to believe back in the days when some argued that the firing of c-fibres was identical with the occurence of pain. Fink wants to draw a distinction between pain itself, a sensation, and suffering, the emotional response associated with it; it is the latter, in his view, which is the bad thing while pain itself is a mere colourless report. As a further argument he notes research which seems to show that when subjects are feeling compassion, some neural activity can be seen in areas which are normally active when the subjects themselves are feeling pain. The subjects, as it were, feel the the pain of others, though obviously without actual nociception.

So is Fink right? I think many people's first reaction might be that unpleasantness just defines pain, so that if you're feeling something that isn't unpleasant, we wouldn't want to call it pain. We might say that people with asymbolia experience nocition (not sure that's really a word but work with me on this) but not pain. Fink would say - he does say - that we ought to listen to what people say. Usage should determine our definition, he says, we should not make our definitions normatively control our usage. But he's in a weak position here. If we are to pay attention to usage, then surely we should pay attention to the usage of the vast majority of people who regard pain as a unitary phenomenon, not to a small group of people with a most unusual set of experiences which might have tutored their perceptions in unreliable ways. I'm not sure it's clear that asymbolics, in any case, insist that what they're aware of is proper, echt pain - if they were asked, would they perhaps agree that it's not pain in quite the ordinary sense?

I'm not convinced that suffering, or unpleasantess, is really a well-defined entity in the way Fink requires. Unpleasantness may be a slight lapse of manners at a tea-party; you might suffer badly on the stock exchange while happily sipping a cocktail on your sun-lounger. I'm not sure there is a distinct complex of emotional affect we can label as suffering at all. And if ther is, we're back with the sheer implauisibility of saying that that's what the bad stuff is: when I hit my thumb with a hammer it doesn't seem like a matter of affect to me, it seems very definitly like old-fashioned simple pain.

If we're going to take that line, though, we have to account for Frink's admittedly persuasive examples, in particular asymbolia. Never mind now what we call it: how is it that these people can experience something they're willing to call pain without minding it,k if it isn't that our concept of pain needs reform?

Well, there is one other property of pain which we've overlooked so far. There is one obvious kind of pain which I can perceive without being disturbed at all - yours. We may indeed feel some sympathetic twinges for the pain of others, but a key point about pain is it's essentially

ours. It sticks to us in a way nothing else does: it's normal in philosophy to speak of the external world, but pain, perhaps uniquely, isn't external in that sense: it's in here with us. That may be why it has another property, noted by Fink, of being very difficult to ignore.

So it may be that subjects with asymbolia are not lacking emotional affect, but rather any sense of ownership. The pain they feel is external, it's not particularly theirs: like Mrs Gradgrind they feel that¶

> '*… there's a pain somewhere in the room, but I couldn't positively say that I have got it.*'

Zoned-out Rat Consciousness

29 October 2011

New Scientist suggests that zoned-out rats may give us a clue to consciousness.

It's all to do with the Default Mode Network, or DMN. You might think that when we stop concentrating on a particular task and sit back for a few quiet minutes the level of activity in our brains would fall, but it turns out this isn't really so: instead, more or less the same level of activity appears to continue, but it switches to a different set of areas - in particular, a linked set of areas in the cortex and elsewhere. This is the DMN, but what is it doing?

A completely honest answer, I think, would be that we don't exactly know except that it's something other than concentrating on a task. In human subjects the DMN seems to be associated with daydreaming, but also with other detached modes of thought. Why would this help explain consciousness? It seems that in patients with locked-in syndrome, where consciousness is fully retained but the patient is unable to move, the DMN is functioning normally, whereas in persistent vegetative syndrome, where consciousness is absent, it is disrupted.

I can think of a further reason to think that this might shed light on consciousness. It's not much of a stretch to see DMN activity as being the kind of thinking that isn't directly related to inputs and outputs. When we're working on a task those are crucial, but one plausible account of the role of consciousness is exactly that it lets us escape from giving instant responses to our surroundings and lets us develop longer-term plans, deeper understanding, and more complex behaviour. If the DMN represents useful mental activity detached from inputs and outputs it is exactly the thing whose existence the behaviourists denied, which is pretty much the same as one conception of consciousness.

The New Scientist and others speak of the DMN as associated with introspection, but I can't see the evidence for that. To be daydreaming or thinking in general terms about stuff that is or might be going on is not introspection. I think there's some confusion going on here between thinking internally and thinking about what's going on internally: and perhaps a further suggestion that introspection= self-awareness = consciousness: those are tenable but debatable equations which don't seem to be vindicated or disproved by the mere existence of the DMN. So perhaps the excitement is premature.

The rats are not that reassuring either. The New Scientist reports that analogues of the human DMN have been found in monkeys, and now even in rats. That's interesting, but unless we rate the consciousness of rats unusually highly it seems to show that the DMN cannot explain any uniquely human level of consciousness. Fair enough: I don't disdain rat consciousness altogether: but it's worse than that because, as I understand it, the evidence currently suggests that younger human children don't have an identifiable DMN. It would be somewhat weird to attribute to rats a level of reflective consciousness which is absent in human infants - wouldn't

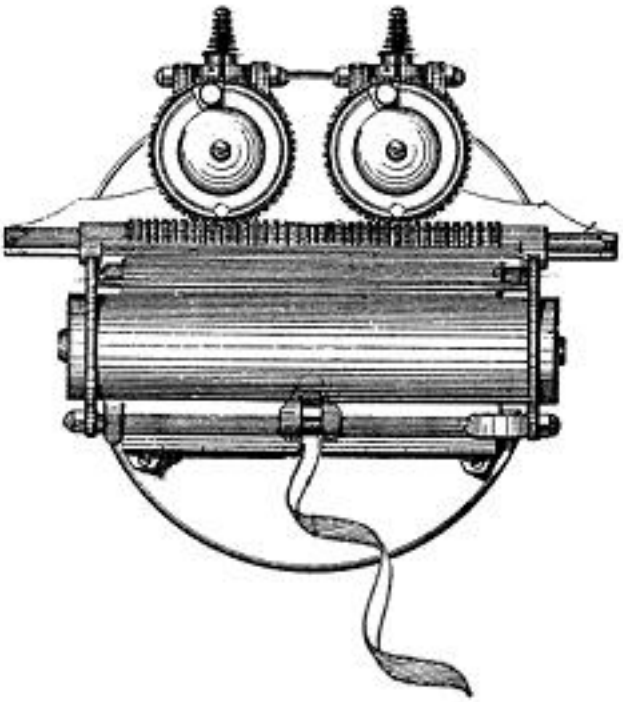
it? If more were needed to put us off, it is not quite 100% agreed that the DMN is in fact a functional entity in itself; it could yet turn out to be more like the mere absence of the TPN, the Task Positive Network which is its opposite (or complement) - the similar set of areas which appear to work together when we're engaged in a specific task. Perhaps the level of neuronal activity in the brain stays high, not because the DMN is really processing anything, but because the brain just uses a lot of energy to tick over?

Still, if the DMN doesn't explain what consciousness is, it's hard to resist the view that it's telling us something about how it works. Problems with the DMN have been put forward as possible causes of Alzheimer's, autism, and schizophrenia (I think everything has been put forward as a possible cause of schizophrenia). The range of problems is perhaps an indication of the vagueness of the theories. There is some good evidence of a correlation between Alzheimer's and disrupted DMN: but then the DMN includes quite a siginficant sampling of some important areas of the brain, so that may not mean all that much. It could be that when consciousness is disrupted the DMN tends naturally to get disrupted too, without that implying that the DMN actually runs or constitutes even the less-focused forms of consciousness.

At the end of the day what we're left with is that our brains - and even rat brains - don't use the same circuits for task-related and non-task related activity, but go through a fairly large-scale switch of resources. Even if we're idly daydreaming about driving into town already, it seems we bring in a different set of neurons to do it with. There has to be some good reason for this, but what...?

Output Consciousness

6 November 2011



The analogy with a digital computer has energised and strongly influenced our thinking about the human mind for at least sixty years, beginning with Turing's seminal paper of 1950, 'Computing machinery and intelligence', and gaining in influence as computers became first real, and then ubiquitous. Whether or not you like the analogy, I think you'd have to concede that it has often set the terms of the discussion over recent decades. Yet we've never got it quite clear, and in some respects we've almost always got it wrong.

In particular, I'd like to suggest: consciousness is an output, not processing.

At first sight it might seem that consciousness can't be an output, on the simple grounds that it isn't, well, put out. Our consciousness is internal, it goes on in our heads - how could that be an output? I don't, of course, mean it's an output in that literal sense of being physically emitted: rather, I mean it's the final product of a process, in this case a mental process. It may often be retained in our heads, but in some sense it's the end of the line, the result.

It may be worth noting in passing that consciousness is pretty strongly linked with outputs in the simpler sense, though: so much so that the Turing test is based entirely on the ability of the testee to output strings of characters which gain the approval of the judges. Quality of output is taken to be the best possible sign of the presence of consciousness.

Wait a minute, you may say, consciousness isn't a final output, it's surely part of the process: what goes on in our conscious mind feeds back into our further thoughts and our behaviour. That's the whole point of it, surely; to allow more complex and detached forms of processing to take place so that our true outputs in behaviour will eventually be better planned and targeted?

It's true that the contents of consciousness may feed back into our mental processes, and that must be at least partly why it exists (its role in forming genuine verbal outputs is probably significant too) - I'm not suggesting consciousness is a mere epiphenomenon, like, as they say, the whistle on a train. Items from consciousness may be inputs as well as outputs. To take an unarguable example, I've never managed to remember how many days there are in each month: but I have managed to remember that little rhyme which contains the information. So if I need to know how many days there are in August, I recall the rhyme and repeat it to myself: in this case the contents of my consciousness are helpfully fed back into my mind. Apart from clunky manoeuvres of this kind, though, I think careful introspection suggests consciousness does not feed directly back into the underlying mental processes all that often. If we want to make a decision we may hold the alternatives in mind and present them to ourselves in sequence, but what we're waiting for is a feeling or a salient piece of reasoning to pop into our minds from

some lower, essentially inscrutable process: we're not normally putting our own thoughts on the subject together by hand. I think Fodor once said he had no conscious access to the mental processes which produced his views on any philosophical issue: if he inspected the contents of his mind while cogitating about a particular problem all he came up with were sub-articulate thoughts approximately like "Come on, Jerry!" I feel much the same.

With apologies if I'm repeating things I've said before, I think it may help if I mention some of the confusions that I think arise from not recognising the output nature of consciousness. A striking example is Dennett's odd view that consciousness might involve a serial computer simulated on a parallel machine. We know, of course, that when people speak of the brain being 'massively parallel' they usually mean that many different functional areas are promiscuously interconnected, something radically different from massively parallel computing in the original sense of a carefully managed set of isolated processes; but Dennett seems to be motivated by an additional misunderstanding in which it is assumed that only a serial process can give rise to a coherent serial consciousness. Not at all: the outputs from parallel and serial processing are identical (they'd better be): it's just that the parallel approach sometimes gets there quicker.

It's a little unfair to single out Dennett: the same assumption that properties of the underlying process must also be properties of the output consciousness can be discerned elsewhere: it's just that Dennett is clearer than most. Another striking example might be Libet's notorious finding that consciousness of a decision arrives some time after the decision itself - but of course it does! The decision is an event in processes of which consciousness is the output.

It's hard to see consciousness as an output, partly because it can also be an input, but also because we identify ourselves with our thoughts. We want to believe that we ourselves enjoy agency, that we have causal effects, and so we're inclined to believe that our thoughts are what does the trick - although we know quite well that when we move our arm it's not thinking about it that makes it happen. This supposed identity of thoughts and self (after all, it's because I think, that I am, isn't it?) is so strong that some, failing to find in their thoughts anything but fleeting bundles of momentary impressions , have concluded there is no self after all. I think that level of scepticism is unwarranted: it's just that our selves remain inscrutably shadowed to direct conscious observation. "Know thyself", said the inscription on the temple of the Delphic oracle - alas, ultimately we can't.

Retinoid Consciousness

25 November 2011

It's not often these days, it seems to me, that a 'theatre of consciousness' is proposed or defended: if the idea is mentioned, it's more commonly to dismiss it or to distance a theory from it. All the more credit then, to Arnold Trehub whose Retinoid theory is just that, and a bold and sweeping example, too.

A retinoid system, at its simplest, is an array of neurons which captures and retains a pattern of activity in the retina, the sensitive layer at the back of the eye which translates light into neuronal activity. The retinoid array retains the topology of the original pattern and because the neurons are autaptic, ie they stimulate themselves, that pattern can be retained for a while. Naturally we need more than one such array to do anything much, and in fact the proposition is that a series of them act together to form a model of three-dimensional space. That makes it sound a simple matter, but of course it isn't. Trehub proposes a series of mechanisms by which raw activity in the retina can be translated into a stable, three-dimensional model, allowing for the way our foveas (the small central parts of the retinas that actually do most of the important work - our vision is actually only sharp in a tiny area of the visual field) sweep restlessly but selectively across the scene and so on.

Those proposals look ingenious and plausible. I wonder a bit about what the firing of a neuron in the retinoid model of space actually represents, in the end? It isn't simply patterns of retinal activity any more, because the original patterns have been amended and extended – so you might argue that the system is not as retinoid as all that. It's tempting to think that each point of activity is a bit like a voxel, a three-dimensional pixel or a single tiny brick in a model world, but I'm not sure that's right either. It certainly goes beyond that because Trehub wants to bring in all the sensory modalities. How a spatial model works in this respect is not fully clear to me – I'm not sure how we deal with the location of smells, for example, something that seems inherently vague in reality. Perhaps smells are always registered in the retinoid as situated within the nose, although that doesn't seem to capture the experience altogether; and what about hearing? It's a common enough experience to hear a sound without having any clear idea of its point of origin, but it won't quite do to say that sounds are located in the head either. I don't know how "sounds sort of like it's coming from over there" is represented in a spacial retinoid model. If we're going on to deal with the whole world of mental phenomena, we're clearly going to run into more problems over location – where are my meditations, my emotions, my intimations of immortality, my poignant apostrophisation of the snows of yesteryear?

Another key element in the system is a self-locus retinoid whose job is essentially to record the presumed location of the self in experienced space (and also, if I've understood correctly, imagined space, because we can use other retinoid structures to model non-existent spaces, or represent to ourselves our moving through locations we don't actually occupy). This provides the basis for our sense of self, but importantly there's much more to it than that.

Trehub suggests that the brain also provides a semantic token system, so that when we see a dog, somewhere in our brain a set of neurons that make up the 'dog' token start firing. We also have a token of the self-locus (roughly speaking this amounts to a token of ourselves or

something like the idea of 'me'), which Trehub designates I! This gives us our sense of identity and pulls together all our tokens and perceptions. It also has an interesting role in the management of our beliefs. In Trehub's view our beliefs are encoded by tokens in a semantic network as explicit propositions: whether we consider these propositions true or false is determined by a linkage with a special token which indicates the truth value, while a similar linkage to I! determines whether these beliefs belong to us. I presume a linkage to the 'Fred' token makes them, in our view, Fred's beliefs, providing the basis of a 'theory of mind'.

We now have in place the essentials for the theatre of consciousness: a retinoid-supported model of egocentric space (Trehub calls it egocentric, though it seems to me the self is always just to the side of it looking in), and a prime actor endowed with beliefs and perceptions. Trehub denies that we need an homuncular audience, a 'little man in our head'. He casts us as actors on our own stage; if there is an audience it's the unconscious parts of the mind. He accepts that the theatre is in a sense illusory – the representations are not the reality they stand in for – but he wishes to keep the baby even at the price of retaining that much bathwater.

So what can we say about all this? As a theory of consciousness it's dominated to an unusual degree by vision and space. This fits Trehub's view of consciousness: he regards the most basic form as a simple sense of presence in a space, with more complex levels involving an awareness of self and then of everything else. I'm not sure a spatially-based conception fits my idea of consciousness so well. When I examine my own mind I find many of the most salient thoughts and feelings have no location; it's not that their location is vague or unknown, space just doesn't come into it. Trehub suggests that when we think of a dog, we're liable to have an image of a dog in our mind: while we can have one, it doesn't seem an essential part of the process to me. If I'm thinking of an arbitrary dog, I don't call up an image, let alone one situated in an imaginary space: if you ask me what colour or breed my arbitrary dog is, I have to make something up in order to answer. (In fact the position is complicated because reflection suggests there are actually at least half-a dozen different ways of thinking about an arbitrary dog: some feature images, some don't.)

Second, the theory centres on a model of space, but I'm not sure how much having a model does for us. Putting it radically, if you can work out what you're going to do with your model, why not do that to reality instead and throw the model away? Now of course, models are useful for working on counterfactuals; predicting outcomes, testing contingencies and plans, and so on, but here we're mainly talking about perception. If we stick to pure perception, doesn't it seem that the model merely introduces another stage where errors might creep in, a stage which apparently insulates us from reality to the point that our perceptions are actually in some sense illusions? Trehub might respond that the model has evolved for its value in dealing with predictions and planning, but also provides current perception; maybe, but there's another problem there. The idea that my current perception of the world and my plans about next week are essentially running on the same system is not intuitively appealing – they feel very different.

The inevitable fear with a model-based system is that the real work is being deferred to an implied homunculus; that the model is, in effect, there for a 'man in our head' to look at. Trehub of course denies this, but the suspicion is reinforced in this case by the way the model preserves the topology of the retinal image: isn't it a little odd that, as it were, the process of perceiving a shape should itself have the same shape?

Trehub has a robust response available, however; the evidence shows clearly that the brain does in fact produce models of perceived objects, filling in the missing bits, resolving

ambiguities and making inferences. Many optical illusions arise from the fact that the pre-conscious parts of our visual system don't tell us which bits of the world they've hypothesised, but present the whole thing as truly and unambiguously out there. Perception is not, after all, just a simple input procedure but a complex combination of top-down and bottom-up processes in which models can have an essential role. And while I don't think the retinoid structure specifically has been confirmed by research, so far as my limited grasp of the neurology goes it seems to be fully consistent with the way things seem to work. Although it may still seem surprising it's undeniably the case, for example, that the visual pathways of the brain preserve retinal topology to a remarkable degree. So as a theory of perception at least, the retinoid system is not easily dismissed. As a model of general consciousness, I'm not so sure. Containing a model of x is not, after all, the same thing as being aware of x. We need more.

What about the self, then? It's natural that given Trehub's spatial perspective he should focus on defining the location of the self, but that only seems to be a small, almost incidental part of our sense of self. Personally, I'm inclined to put the thoughts first, and then identify myself as their origin; I identify myself not by location but by a kind of backward extrapolation to the abstract-seeming origin of my mental activity. This has nothing to do with physical space. Of course Trehub's system has more to it than mere location, in the special tokens used to signify belonging to me and truth. But this part of the theory seems especially problematic. Why should simply flagging a spatial position and some propositions as mine endow a set of neurons with a sense of selfhood, any more than flagging them as Fred's? I can easily imagine that location and the same set of propositions being someone else's, or no-one's. I think Trehub means that linking up the tokens in this way causes me to view that location as mine and those propositions as my beliefs, but notice that in saying that I'm smuggling in a self who has views about things and a capacity for ownership; I've inadvertently and unconsciously brought in that wretched homunculus after all. For that matter, why would flagging a proposition as a belief turn it into one? I can flag up propositions in various ways on a piece of paper without making them come to intentional life. To believe something you have to mean it, and unfortunately no-one really knows what 'meaning it' means – that's one of the things to be explained by a full-blown theory of consciousness.

Moreover, the system of tokens and beliefs encoded in explicit propositions seems fatally vulnerable to the wider form of the frame problem. We actually have an infinite number of background beliefs (Julius Caesar never wore a top hat) which we've never stated explicitly but which we draw on readily, instantly, without having to do any thinking, when they become relevant (This play is supposed to be in authentic costume!): but even if we had a finite set of propositions to deal with the task of updating them and drawing inferences from them rapidly becomes impossible through a kind of combinatorial explosion. (If this is unfamiliar stuff, I recommend Dennett's seminal cognitive wheels paper.) It just doesn't seem likely nowadays that logical processing of explicit propositions is really what underlies mental activity.

Some important reservations then, but it's important not to criticise Trehub's approach for failing to be a panacea or providing all the answers on consciousness – that's not really what we're being offered. If we take consciousness to mean awareness, the retinoid system offers some elegant and plausible mechanisms. It might yet be that the theatre deserves another visit.

[I was flattered to discover an extract from this piece being used in a writing course as an example of 'polite scepticism'.]

Smelling Secret Harmonies

18 December 2011

Is trilled smell possible? Ed Cooke and Erik Myin raise the question in the JCS. Why do we care? Well, for one thing smell has always tended to be the poor relation in discussions of conscious experience. The science of vision is so much better developed that seeing generally looks a more tractable area to attack, but arguably the discussion is somewhat lop-sided as a result; 'seeing red' isn't necessarily a perfect epitome of all sensory experience, so a bit of clarification around smells might well be useful.

But the main point of asking the question is to test what Cooke and Myin call the independence thesis: the view that the experienced character of sensations includes a 'something it is like' over and above the gross physics of the business: that there's an ultimate smelliness about smell that has nothing to do with the details of the sensory process. I would say there's a range of possible positions here. Hardly anyone, I think would say that the physics of perception is irrelevant to how the experience seems. We know that that the wave structure of light and sound determines some of the characteristics of the experiences of vision and hearing, for example, and we know that smell is vaguer about location than vision because it depends on gases wafting around rather than sharply defined rays of light. But beyond that the consensus breaks down. Some would say that these physical characteristics are just the basics and the real excitement lies in the ineffable qualities of experience. Purple is a thing in itself, not a blank sensory token which would do equally well for the smell of coffee, they might say.

Some would go further and accept that the qualities of experience are very largely determined by the physics of the medium and sensory apparatus, but that there's a certain something beyond that which doesn't reduce to the simple physics. Rigorous materialists will be tempted to go further still and take the view that however complex and indefinable our experiences may seem, they are fully determined by the qualities of the processes that give rise to them: this of course, amounts to denying the existence of ineffable qualia. (My own view, for what it's worth, lies an infinitesimal distance short of this extreme.)

Cooke and Myin's approach is to look at the consequences of the independence thesis. If it's true then we ought to be able to transfer the forms of one sensory modality to another without it losing its identity. So, in sound we can have a trill, a very rapid alternation of two notes; if independence is true, we ought to be able to have trilled smells.

Before tackling the thought experiment in more detail, Cooke and Myin provide a brisk review of some of the relevant science, including some odd and interesting facts. The smell of pressure-cooked pork liver is made up of 179 different compounds; airflow is indispensable to smell (having your nose full of smelly stuff or your receptors stimulated produces no sensation unless there's airflow); and so sniffing is more important than you may have thought. It turns out that human beings are pretty well incapable of identifying single components of a smell when there are more than three - so much for perfume designers - and perception of smell is also very heavily conditioned by previous experience (if you've encountered smell b together with lemony smells in the past, you'll tend to think smell b has lemony notes even when the lemon smells are objectively absent). It looks as if we might each be working with a typical vocabulary of about 10,000 known smells, out of a theoretical 400,00 that the nose can distinguish: best

estimates suggest that smell-space has a minimum of somewhere between 32 and 68 dimensions (as compared to human colour vision's paltry 3).

Now we come to the thought-experiment itself. It seems that Jesse Prinz has denied the possibility that a sound could become a smell merely by changing the structure of the experience (could the sound of a fire alarm ever become the smell of smoke?), so with fine daring, that is the first transition Cooke and Myin propose to anatomise in a thought-experiment.

Thought-experiments are always a little unsatisfactory because they don't really force people to accept your conclusions in the way that a proper argument does. In this case, moreover, it seems to me there's a particularly difficult trick to bring off because for the experiment to convince, Cooke and Myin want the transfer of properties to seem plausible; yet the more plausible it seems the more plausible independence seems too. They want us to believe that they're doing the best possible description of a transfer that could plausibly happen, in order to convince us that once we understand it it's not plausible that it's really a transfer at all.

However, I think they do a commendable job. First, sounds have to become less distinct in their onset and direction; they have to be more like generalised hums which float around appearing and dispersing slowly (no good for rapid warnings any more). The we have to imagine that we use our noses to detect sounds, that they only become perceptible when we breathe, and that sniffing or breathing deeply affects their intensity. We must imagine that it's now a little more difficult to pick out single sounds when there are several at once: we might have to think about it for few moments and take some extra sniffs.

That's not too bad, but there are bigger problems. We've noted that smell space appears to be huge; Cooke and Myin suggest we could enlarge sound space the same way by imagining that the differences in sound are like the differnces in timbre between musical instruments (though we have to suppose that we can readily distinguish the timbres of 10,000 or so different instruments). On the other hand, musical notes fit on an organised scale with perceptible relationships between different notes: smell doesn't really have that, so we must drop it and assume that sounds are essentially monotonous. To round things off with behavioural factors, we should think of sound as no longer used for communication, but mainly for the evaluation of the acceptability of food, people and other biological entities; and we should imagine that sounds now have that characteristic of certain smells which allows them to evoke memories with particular potency.

If you're still with the experiment, you'll now have some intuitive idea of what it would be like if sounds had the structural and other characteristics of smells. But no, say Cooke and Myin: isn't it apparent now that the sensations we're talking about wouldn't be sounds any more (in fact they would pretty much have become smells)? Isn't it clear, in short, that in order to be trillable, smells would have to cease being smells? They go on to a further thought experiment in which smells become colours.

This is a valuable exercise, but as I say, thought experiments are not knock-down arguments, and I am willing to bet there wil in fact be plenty of people who are prepared to go along with Cooke and Myin's transition but insist at the end that the sensations they're imagining are still in some way sounds, or at least have a core soundiness which makes them different from echt smells. (You notice how I criticise the weakness of thought experiments and yet here I am doing

something worse - a kind of third-person thought-experiment where I invite agreement that in certain odd circumstances other people would think in a certain way.)

Personally I think some of the most interesting territory revealed here is not so much at the ends of the transition as in the middle. The experiment raises the possibility of mixed modalities never before imagined, chimerical experiences with some of the characteristics of two or more different standard senses. Not just that, either, because we can invent new physical constraints and structures and develop possible sensory modalities which have nothing whatever in common with any real ones, if our imagination permits.

This gives Cooke and Myin some possible new ammunition. Do all these imaginary new modalities get their own essence, their own qualia? If we mix smell and hearing in different ways, do we have to suppose that there are distinct qualia of, er, smearing and hell?

For that matter, what if we took a subject (all right, victim) through the transition of sound to smell; and then separately gave him back sound? has he now got two distinct experiences of sound? Then if we move the new sound2 though the transition to smell, does he have two smells? And if we then give him back a separate sense of sound again? And so on.

I can't help thinking it would be quite a Christmas present if we could have a sense with the spatial distinctness of vision, the structured harmonics of sound, and the immense dimensionality of smell. There would be some truly amazing symphonic odours to be painted.

The Pongid Theory of Mind

14 January 2012

Researchers from the Max Planck Institute and St Andrew's University have come up with some fresh evidence that chimps have a theory of mind (ToM) – that is to say that they are aware that other individuals possess knowledge and that what they know doesn't always match what we know.

The researchers placed dummy snakes in the path of wild chimps: the chimps gave warning calls more frequently in the presence of others who, so far as they could tell, had no prior knowledge of the presumed hazard.

This kind of research is fraught with difficulty. Morgan's Canon tells us that we should only use consciousness as an explanation for some item of behaviour where no simpler explanation is available, and similarly we should be reluctant to grant chimps ToM unless there is no alternative. Couldn't the explanation be, for example, that chimps who are alone are more likely to give warning calls, either because that response is just hard-wired, or because they are more fearful when alone? Alternatively, perhaps the observed behaviour could be largely explained if chimps are programmed to give a warning call, but only one, for each member of the troupe they spot or hear approaching?

Although I think Morgan's Canon is absolutely the right kind of principle to apply, it is difficult to satisfy, and if read too literally perhaps impossible. We know from all the discussions of philosophical zombies that there are plenty of thoughtful people who find it conceivable that all of human behaviour could be produced without consciousness (at any rate, without the kind of consciousness that requires actual phenomenal subjective experience). If that's really true then there are surely no cases in which behaviour can strictly be explained only by consciousness. It's equally hard, going on impossible, to rule out every conceivable alternative explanation for the chimps' behaviour – but the researchers were well aware of the problem and the key point of the research is the observation of circumstances where, for example, chimp A could be presumed to have heard an earlier warning, but chimp B could not. So we can take the claims they make as well grounded. It seems that with some inevitable margin of doubt we can reasonably take it as established that chimps do have ToM.

So what? We might have been willing to assume that that was probably the case anyway. We already know chimps are extremely bright and there are many who believe they can develop language skills which approach human levels. Language is what makes it so much easier to know for sure that human beings have ToM – they can tell us about it – so if chimps are anywhere near that level it's really no surprise that they also have ToM. (Interesting, by the way that the current research uses the chimps' proto-linguistic warning calls.) One further conclusion offered by the researchers themselves is that ToM must have emerged in the primate lineage at a point before the divergence of chimp and human ancestors: but that ain't necessarily so. It could equally be that each lineage has developed a functionally comparable capacity in parallel, one which the latest shared ancestor need never have had.

Do we and our pongid cousins have the same ToM? In some respects obviously not. For one thing, we humans really do have actual academic theories of mind; and we write novels filled with the putative contents of minds that never existed. We have ToM on levels which completely transcend the mental lives of chimps. Are these, though, just fancy overlays on an underlying ability which remains essentially the same? Alas, there's no easy way of telling without knowing what's going on in the chimp's mind – what it is like to be a chimp – and Nagel long ago told us that that was impossible.

Attempting to know the unknowable is nothing new for us, though, so let's at least briefly try to achieve the impossible. There are lots of possibilities for what might be passing through the chimp's mind: by way of illustration, it could be any of the following.

1. A cloudy sense of something indefinable but importantly snake-related which is missing in Chimp B.
2. A mental picture of Chimp B continuing to advance and stumbling on the snake.
3. A brief empathetic sense of being Chimp B, and a recollection that seeing the snake or hearing a warning has not occurred.
4. Routine enumeration of the troupe and its whereabouts leading to a realisation that Chimp B hasn't been around for a while.
5. Occurrence of proto-verbal content equivalent to uttering the sentence "Look there's B, who doesn't know about the snake yet!"

There are plenty of other possibilities: cataloguing them would in itself be a challenging task. Moreover, humans are clearly capable of operating on two or more of these levels at once, and it would be mere speciesism to assume that chimps are not. Still, can we pare it down a bit: given that chimps lack full-blown human linguistic abilities and are relatively limited in their foresight, can we plausibly hypothesise that cases like 5 and other involving relatively complex levels of abstraction are probably absent from the chimp experience? I'm not sure, and even if we can it doesn't help all that much.

So instead I ask myself what state obtained in my own mind last time I warned someone about a potential hazard. Luckily I do remember a couple of occasions, but interestingly introspection leaves me quite uncertain about the answer. This could be a result of hazy memory, but I think it's worse than that: I think the main problem is that so far as conscious thought goes I could have been thinking anything. It feels as if there is no distinct single state of mind which corresponds to noticing that somebody needs to be warned about something; curiously I feel tempted to examine my own behaviour and conclude that if I did go on to warn someone, I must have been thinking that they needed warning.

That kind of approach is another option, I suppose: we can take a behaviourist tack and say that if chimps behave in a way that displays ToM, then they have it, and that's all there is to be said about it. If we can't formulate clearly what kind of behaviour that would be, that just means ToM itself turns out to be mentalistic nonsense. The snag with that is that ToM is pretty certainly mentalistic nonsense to behaviourists anyway; so if we think the question is worth answering we have to look elsewhere.

We could get neuronal on this: we might, for example, be able to scan human and chimp brains and detect some distinctive patterns of activity which occur just when the relevant primate appears to be getting ready to issue a warning. If these patterns of activity occurred in the corresponding sections of the chimp and human brain (perhaps involving some of those special mirror neurons) we should be inclined to conclude that our ToMs were basically the same: if

they occurred in different places we should be very tempted to conclude that evolution had recruited different sections of the two species' brains to carry out the same function. This latter case is quite plausible – in human brains, for example, the areas used for speech don't match the bits of the chimp brain used for vocalisations (which apparently correspond to areas used by humans only for involuntary gasps and cries and, strangely enough, for swearing).

Results like that might settle the evolutionary question; but not the deeper philosophical one. Even if we did use a different set of neurons, it wouldn't prove we weren't running the same ToM. Different human beings certainly use somewhat different arrays of neurons – no two brains are wired identically. If we came across the yeti and found he was fully up to human levels of consciousness, able to hold an impeccably normal human-style conversation with us and discuss ToM just as we do, and then we made the astonishing discovery that he had no prefrontal cortex and was using what in humans would have been his cerebellum to do his conscious thinking with, we would not on that account alone say he had a different kind of consciousness (at least, I don't think we would).

So it looks to me as if we have a radical pattern of variation at both ends. All sorts of neuronal wiring (or maybe silicon or beer cans and string – why not?) will do at the bottom level; all sorts of cogitative content will do at the top levels. Somewhere in the middle is there a level of description where deciding that someone needs to be warned is just that and nothing else, and where we can meaningfully compare and contrast human and chimp? I suspect there is, but I also suspect that it resides in something analogous to a high-level mental metacode of a kind we should need a proper theory of mind even to begin imagining.

Crazyism

29 January 2012

Eric Schwitzgebel has done a TEDx talk setting out his doctrine of crazyism.

The claim, briefly is that any successful theory of consciousness - in fact, any metaphysical theory - will have to include some core elements that are clearly crazy. “Crazy” is taken to describe any thesis which conflicts with common sense, and which we have no strong epistemic reason for accepting.

Schwitzgebel defends this thesis mainly by surveying the range of options available in philosophy of mind and pointing out that all of them - even those which set out to be pragmatic or commonsensical - imply propositions which are demonstrably crazy. I think he’s right about this; I’ve observed myself in the past that for any given theory of consciousness the strongest arguments are always the ones against: unnervingly this somehow remains true even when one theory is essentially just the negation of another theory. Schwitzgebel suggests we should never think that our own preferred view is, on balance, more likely than the combination of the alternatives - we should always give less than 50% credence to our preferred view or, if you like, never quite believe anything.

I won’t recapitulate Schwitzgebel’s case here, but it did provoke me to wonder about the issue of what we would find acceptable as an answer to the problem of consciousness. It’s certainly true that some theories would not, as it were, be crazy enough to appeal. Suppose the Blue Brain project triumphed, and delivered a brain simulated down to neuronal level. We could run the simulation and predict the brain’s behaviour; for anything the simulated person said or thought, we could give a complete neuronal specification, and in that sense a complete explanation. But it wouldn’t seem that that really answered any of the deeper questions.

Equally, for all those theories that tell us there’s really nothing to explain, our consciousness and our selfhood are just delusions generated by aspects of the mental mechanism, one problem is that the answer seems too easy (though in another sense these views are surely crazy enough). We don’t want to be told to move along, nothing to see here, folks; what we want is an “Aha!” moment, a theory that makes things suddenly fall into places where they make dramatic new sense. How do we get such moments?

I think they come from a translation or bridge that lets us see how one understood realm transfers across into another realm which is also understood but not connected. Maybe we find out that Hesperus is Phosphorus and that both are in fact the planet Venus; then the strange behaviour of the evening star and the morning star suddenly make more sense. Another related way of generating the Aha! is to discover that we have been conceptualising things wrongly: that two things we thought were separate are really aspects of the same thing, or that a thing we took to be a single phenomenon is actually two different things we have conflated together: temperature and heat, for example.

It certainly looks as if consciousness is right for an Aha! of those kinds - we have the two separate realms, the mental and the physical, all ready to go. But Colin McGinn has argues that

the very distinctness of the realms means that no explanation can ever be forthcoming, and many people since Brentano have shared the sense that no bridging or reshuffling of concepts is even conceivable. The thing is, we don't get the kind of paradigm shift we need by labouring away within the existing framework: we need something to jolt us out of it and there's no telling what. We know now that in order to come up with the theory that speciation occurs through differential survival of the fittest, you needed to visit the tropics and collect a lot of examples of local fauna, read Malthus and then fall ill. Darwin and Wallace both had their dogmatic slumbers shaken up in this way; but it was not evident in advance that that was what it took. Perhaps even now a young doctor who has treated schizophrenics, has had the required motorbike accident and is just about to read the text on encryption which is an essential precursor to the Theory.

I sort of hope and believe that something like that is the case and that when the Theory is available we shall see that one of those theses that look crazy now are not quite what we thought: so crazyism will turn out to be false, or at any rate only provisional. But Schwitzgebel's essential pessimism could turn out to be justified. We could end up with a theory like quantum mechanics, which seems to do the job so far as anyone can see, but which just refuses to click with our brains positively enough for the Aha! moment.

Schwitzgebel doesn't spend much time on the wider claim that metaphysics as a whole is crazy, but it's an interesting possibility that the problem doesn't really lie with philosophy of mind but something altogether deeper. Maybe we need to look away from the mind as such and spend some time on… what? Causality? Basic ontology? Once again, I have no idea.

So I Hear

1 February 2012

We have become accustomed over the years to exaggerated claims about brain research. I think my favourite will always be the unexpected claim from British Telecom in 1996 that they were to develop a chip, by 2025, which would fit behind the human eye. I suspect the original idea was to record all retinal activity and hence have, in a sense, a record of the person's whole visual experience – an extremely ambitious but not inherently impossible goal; but somewhere along the way they convinced themselves they would be able to record, not just optical input, but people's thoughts, too. Realising there was no point in under-selling this amazing future achievement, they announced they were going to call it the 'Soul Catcher'. Dr Chris Winter was prepared to go still further and went on record as saying that this was actually "the end of death". I wonder how the project is getting along now?

Not many researchers can match the scope of Dr Winter's imagination or the sheer insolence of his chutzpah, but we have often seen claims to have decoded the mind which are essentially based on identifying simple correspondences between scan results and an item of mental activity. The subject is shown a picture of, say, John Malkovich and scan results obtained: then from scan results the researchers succeed in telling with statistically significant rates of success the occasions when the subject is looking at the same picture of John Malkovich and not one of John Cusack. Voila! The secret language of the brain is cracked! It is not shown that similar scan patterns can be obtained from other subjects looking at the picture of John Malkovich, or from the same subject looking at other pictures of John Malkovich, or thinking about John Malkovich, or even from the same subject looking at the same picture the next day. No general encoding of mental activity is revealed, no general encoding of visual activity, in fact we don't even get for sure a general encoding of that particular picture of John Malkovich in that particular subject on that particular day. The only truth securely revealed is that if you have an experience and then soon afterwards another one just like it, you probably use quite a few of the same neurons in responding to it.

So it was with a certain sinking feeling that I heard the BBC announce that researchers had decoded the language of the brain. The radio report was quite definite about it; they could now reconstruct with their receiving equipment words that subjects were merely thinking: they played back the sound of a suitably robotic-sounding word apparently picked up from someone's inner thoughts. They suggested this could be brought into use in identifying and communicating with 'locked-in' patients, those who though immobilised remain mentally alert. Similar reports appear elsewhere in the press today.

The paper behind all this is far more circumspect than the publicity. It makes generally modest and well-supported claims; only in one place does it venture a little speculation, and doesn't pretend then that it's anything else.

What actually happened is that the experimenters took advantage of an unusual therapeutic situation which allowed them to record directly from electrodes on the brain – a technique which yields far better resolution than any form of scanning. They read their subjects a list of words and noted patterns of activity; they were then able to produce a programme which

automatically reconstructed the characteristics of the sound being heard, sufficiently well for the right word from the list to be identified with a high level of success.

This is not without interest – it sheds some light on the brain's processing of heard information. It shows that quite a lot of information about the actual sound survives at least some distance into the processing – a result we can perhaps compare with visual processing. The kind of worries I alluded to above about generalisability are not absent here, but we do seem, if I've understood correctly, to have got results that should be transferrable and reproducible between subjects.

But reading thoughts? Let's not worry about that for a moment and ask ourselves whether a much-improved version of this technology could tell us what someone was saying as well as what someone was hearing. It seems to me that that would be a whole new game. When the brain interprets sounds as words it necessarily concerns itself with the properties of sound, because that's what it has to deal with; when it delivers an utterance it has no direct interest in the matter, only in tongue, palate, breathing and so on (that may be over-simplifying a touch, admittedly – it would be surprising, for example if feedback didn't play a significant role). It's not likely that the neural patterns for recognising a spoken word are the same as those for speaking it, any more than the neural activity required for reading a word is generally the same as writing it, except inasmuch as both probably involve thinking about the word. For once Heraclitus is wrong: the path up is not the same as the path down.

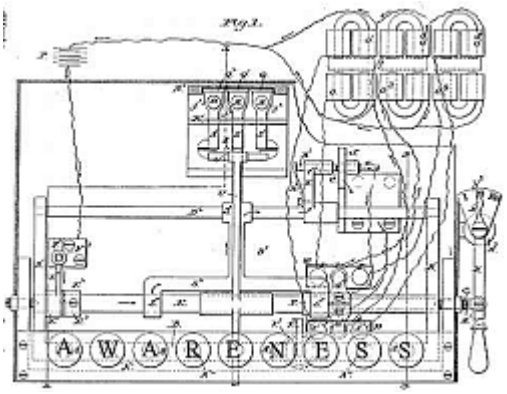
So what about thinking? Is it like hearing, or like speaking? Well, I doubt very much whether thinking of John Malkovich, or even thinking of the words 'John Malkovich' necessarily resembles decoding a sound or preparing to manipulate the lips – unless we are deliberately going through the act of mentally entertaining the idea of hearing or speaking. In the latter case it might plausibly be the case that there are at least some mirror neurons involved in both activities which would bridge the gap between thought and act sufficiently to produce some recognisable activity.

So, if these results can be generalised to a system capable of recognising words in general, and if it's one that demonstrably works for different subjects, and if a way can be found of running it without taking the top of the skull off, and if it turns out that thinking about the sound of a word is sufficiently connected to actually hearing it to stimulate neural patterns which are sufficiently similar that the system can still pick them up in an identifiable form, and if we're talking about someone who is deliberately thinking about the sound of a word, then yes there is some hope here that in that sense we might be able in practice to identify the word being thought.

I suppose that must be worth half a cheer.

Gofai's Not Dead

19 February 2012



I was brought up short by this bold assertion recently:¶

> *… since last two decades, research towards making machines conscious has gained great momentum*

¶ Surely not? My perception of the state of play is quite different, at any rate; it seems to me that the optimism and energy which was evident in the quest for machine consciousness twenty or perhaps thirty years ago has gradually ebbed. Not that anyone has really changed their minds on the ultimate feasibility of a conscious artefact, but expectations about when it might be achieved have generally been slipped back further into the future. Effort has been redirected towards related but more tractable problems, ones that typically move us on in useful ways but do not attempt to crack the big challenge in a single bite. You could say that the grand project was in a state of decline, though I think it would be kinder and perhaps more accurate to say it has moved into a period of greater maturity.

That optimistic aside was from Starzyk and Prasad, at the start of their paper A Computational Model of Machine Consciousness in the International Journal of Machine Consciousness. It's not really fair to describe what they're offering as GOFAI (Good Old-Fashioned Artificial Intelligence) in the original derogatory sense of the term, but I thought their paper had a definite whiff of the heady optimism of those earlier days.

Notwithstanding that optimism their view is that there are still some bits of foundational work to be done…¶

> *there is a lack of a physical description of consciousness that could become a foundation for such models.*

¶…and of course they propose to fill the gap.

They start off with a tip of the hat to five axioms proposed by Aleksander and Dunmall back in 2003 which suggest essential functional properties of a conscious system: embodiment and situatedness, episodic memory, attention, goals and motivation, and emotions. These are said to be necessary but not sufficient conditions for consciousness. If we were in an argumentative mood, I think we could put together a case that some of these are not necessarily essential but we can take it as a useful list for all practical purposes, which is very much where Starzyk and Prasad are coming from.

They give a quick review to some scientific theories (oddly John Searle gets a mention here) and then some philosophical ones. This is avowedly not a comprehensive survey, nor a very lengthy one and the selection is idiosyncratic. Prominent mentions go to Rosenthal's HOT (Higher Order Theory – roughly the view that a thought is conscious if there is another thought about it) and Baars' Global Workspace (which must be the most-cited theoretical model by some distance).

In an unexpected and refreshing leap away from the Anglo-Saxon tradition they also cite Nisargadatta, a leading exponent of the Advaita tradition on the distinction between awareness and consciousness. The two words have been recruited to deal with a lot of different

distinctions over the years: generally I think 'awareness' tends to get used for mere sensory operation or something of that kind which 'consciousness' is reserved for something more ruminative. Nisargadatta seems to be on a different tack:¶

When he uses the term "consciousness", he seems to equate that term with the knowledge of "I Am". On the other hand, when he talks about "awareness", he points to the absolute, something altogether beyond consciousness, which exists non-dualistically irrespective of the presence or absence of consciousness. Thus, according to him, awareness comes first and it exists always. Consciousness can appear and disappear, but awareness always remains.

¶But then we're also told:¶

> *A plant may be aware of a damage done to it [!], but it cannot be conscious about it, since it is not intelligent. In a similar way a worm cut in half is aware and may even feel the pain but it is not conscious about its own body being destroyed*.

¶To allow awareness to a plant seems like the last extreme of generosity in this area - right at the final point before we tip over into outright panpsychism- but in general these remarks makes it sound more as if we're dealing with something like the kind of distinction we're most used to.

Then we have a section on evolution and the development of consciousness, including a table. While I understand that Starzyk and Prasad are keen to situate their account securely in an evolutionary context, it does seem premature to me to discuss the evolution of consciousness before you've defined it. In the case of giraffe's necks, say, there may not be too much of a difficulty, but consciousness is a slippery concept and there is always a danger that you end up discussing something different from, and simpler than, what you first had in mind.

So on to the definition.

> *...our aim was not to remove ambiguity from philosophers' discussion about intelligence or various types of intelligence, but to describe mechanisms and the minimum requirements for the machine to be considered intelligent. In a similar effort, we will try to define machine consciousness in functional terms, such that once a machine satisfies this definition, it is conscious, disregarding the level or form of consciousness it may possess*.

I felt a slight sense of foreboding at this, but here at last is the definition itself.

> *A machine is conscious if besides the required mechanisms for perception, action, learning and associative memory, it has a central executive that controls all the processes (conscious or subconscious) of the machine...*
>
> *...[the] central executive, by relating cognitive experience to internal motivations and plans, creates self-awareness and conscious states of mind*.

I'm afraid this too reminds me of the old days when we were routinely presented with accounts that described a possible architecture without ever giving us any explanation of how this particular arrangement of functions gave rise to the magic inner experience of consciousness itself. We ask ourselves: could we put together a machine with a central executive of this kind that wasn't conscious at all? It's hard to see why not.

I suppose the central executive idea is why Starzyk and Prasad feel keen on the Global Workspace, though I think it would be wrong to take that as genuinely similar; rather than a

central control module, it's a space where miscellaneous functions can co-operate in an anarchic but productive manner.

Starzyk and Prasad go on to flesh out the model, which consists of three main functional blocks - Sensory-motor, Episodic Memory and Learning, and Central Executive; but none of this helped me with the basic question. There's also an interesting suggestion that attention switching is analogous to the saccades performed by our eyes. These are instant ballistic movements of the eye's direction towards things of potential interest (like a flash of light in a dark place); I'm not precisely sure how the analogy works but it seems thought-provoking at least.

They also provide a very handy comparison with a number of other models – a useful section, but it sort of undermines the earlier assertion that there is a gap in this area waiting for attention.

Overall, it turns out that the sense of older times we got at the beginning of the paper is sort of borne out by the conclusion. Many technically inclined people have always been impatient with the philosophical whiffling and keen to get on with building the machine instead, confident that the theoretical insights would bloom from a soil fertilised with practical achievements. But by now we can surely say we've tried that, and it didn't work. Researchers forging ahead without a clear philosophy ran into the frame problem, or simple under-performance, or the inability of their machines to deal with intentionality and meaning. They either abandoned the attempt, adopted more modest goals, or got sidetracked into vast enabling projects such as the compilation of a total encyclopaedia to embody background knowledge or the neuronal modelling of an entire brain. In a few cases brute force approaches actually ended up yielding some practical results, but never touched the core issue of consciousness.

I don't quite know whether it's sort of depressing that effort is still being devoted to essentially doomed avenues, or sort of heartening that optimism never quite dies.

Crimes Against Consciousness

4 March 2012

The Nuffield Council on bioethics is running a consultation on the ethics of new brain technologies: specifically they mention neurostimulation and neural stem cell therapy. Neurostimulation includes transcranial magnetic stimulation (TMS) which typically requires nothing more than putting on a special cap or set of electrodes; and deep brain stimulation (DBS) where the electrodes are surgically inserted into the brain.

All of these are existing technologies which are already in use to varying degrees, and the consultation is prudently geared towards gathering real experience. But of course we can range a bit more freely than that, and it raises an interesting general question: what new crimes can we now commit?

Disappointingly it actually seems that there aren't many really new neurocrimes; most of the candidates turn out to be variations or extensions of the old ones. Even where there is an element of novelty there's often a strong analogy which allows us to transpose an existing moral framework to the new conditions (not that that necessarily means that there are easy or uncontroversial answers to the questions, of course).

I think I've said before, for example, that TMS seems to hold out the prospect of something analogous to the trade in illicit drugs. An unscrupulous neurologist could surely sell wonderful experiences produced by neural stimulation and might well be able to create a dependency which could be exploited for money and general blackmail. The main difference here is that the crucial lever is control of the technology rather than control of the substance, but that is a relatively small matter which has some blurry edges anyway.

It's possible the new technologies might also be able to enhance your brain - if they allow better concentration or recall of information, for example. There is apparently some evidence that TMS might be capable of improving your exam scores. That clearly opens up a question as to whether enhanced performance in an exam, produced by neural stimulation, is cheating; and the wider question of whether easier access to TMS by wealthier citizens would build in a politically unacceptable advantage for those who are already privileged. So far as I know there's no current drug or regime which automatically and reliably boosts academic performance; nevertheless, the issues are essentially the same as those which arise in the case of various other forms of exam cheating, or over access to superior educational facilities. There may be a new aspect to the problem here in that traditional approaches generally rest on the idea that each person has a genuine inherent level of ability; this may become less clear. If a quick shot of TMS through the skull boosts your performance for the next hour only, we might see things one way; whereas if wearing a set of electrodes helps you study and acquire permanently better understanding, we might be more inclined to think it is legitimate in at least some respects. Moreover a boost which can be represented as therapeutic, correcting a deficit rather than providing an enhancement, is far more likely to be deemed acceptable. All in all, we haven't got anything much more than new twists on existing questions.

There is likely to be some scope for improperly influencing the behaviour of others through neural techniques, but this has clear parallels in hypnotism, confidence trickery, and other

persuasive techniques; again there's nothing completely novel here. Indeed, it could be argued that many con tricks and feats of conjuring rest on exploiting neurological quirks as it is.

To steal information from someone's brain is morally not fundamentally different from stealing it out of their diary; and to injure someone by destroying a mental faculty broadly resembles physical injury - the two may indeed go together in many cases.

So what is new? I think if there is fresh scope for evil-doing it is probably to be found in the manipulation of personality and identity. Even here the path is not untrodden, with a substantial history of attempts to modify the personality through drugs or leucotomy; but there is now at least a prospect, albeit still some way off, of far better and more precise tools. As with cosmetic surgery, we might expect the modification of personality to be limited to cases where it has a therapeutic value, together with a range of elective cases over which there might be some argument. The novel thing here is that many cases would require consent; but unlike a nose job, personality modification attacks the basis of consent.

Consider an absurd example in which subject A seeks modification to achieve greater courage and maturity; having achieved both, the improved A now disapproves of the idea of personality modification and insists the changes constitute an injury which must be reversed; once they are reversed, A, with the old personality, wants them done again.

It could be worse; since personality and identity are linked, the new A might take a different line and insist that the changes made in the brain he inhabits were effectively the murder of an older self. This would be as bad a crime as old-fashioned killing, but now it's no good reversing the changes because that amounts to a further murder, and it could be argued that the restored A is in fact not the original come back, but a third person who merely resembles the original; a kind of belated twin. A's brother might sue all the new personalities on the basis that none of them has any more rights to the property and body of the original than a squatter in someone's house.

In circumstances like these there might be a lobby for the view that personality modification should be subject to a blanket ban, in rather the same way that society generally bans us from editing out undesirable personalities with a gun - even our own.

Of course there is in principle another novel crime we might be able to commit: the removal of someone's qualia, their inward subjective experience. This has often been contemplated in the philosophical literature (it is remarkable how many of the most popular thought-experiments in philosophy of mind - whose devotees generally seem the mildest and most enlightened of people - involve atrocious crimes); perhaps now it can become real. The crime would be undetectable since the ineffable qualities it removes could never be mentioned by the victim; the snag is that since there could be no way of measuring our success it's probably impossible to devise or test the required diabolical apparatus...

Transferring Consciousness

26 March 2012

Homo Artificialis had a post at the beginning of the month about Dmitry Itskov and his three-part project:¶

- the development of a functional humanoid synthetic body manipulated through an effective brain-machine interface
- the development of such a body, but including a life-support system for a human brain, so that the synthetic body can replace an existing organic one, and
- the mapping of human consciousness such that it, rather than the physical brain, can be housed in the synthetic body

¶

The daunting gradient in this progression is all too obvious. The first step is something that hasn't been done but looks to be pretty much within the reach of current technology. The second step is something we have a broad idea of how to do - but it goes well beyond current capacity and may therefore turn out to be impossible even in the long run. The third step is one where we don't even understand the goal properly or know whether the ambitious words "mapping of human consciousness" even denote something intelligible.

The idea of transferring your consciousness, to another person or to a machine, is often raised, but there isn't always much discussion of the exact nature of the thing to be transferred. Generally, I suppose, the transferists reckon that consciousness arises from a physical substrate and that if we transfer the relevant properties of that substrate the consciousness will necessarily go with it. That may very well be true, but the devil is in that 'relevant'. If early inventors had tried to develop a flying machine by building in the relevant properties of birds, they would probably have gone long on feathers and flapping.

At least we could deal with feathers and flapping, but with consciousness it's hard even to go wrong creatively because two of the leading features of the phenomenon as we now generally see it are simply not understood at all. Qualia have no physical causality and are undetectable; and there is no generally accepted theory of intentionality, meaningfulness.

But let's not despair too easily. Perhaps we shouldn't get too hooked up on the problems of consciousness here because the thing we're looking to transfer is not consciousness per se but a consciousness. What we're really talking about is personal identity. There's a large philosophical literature about that subject, which was well established centuries before the issues relating to consciousness came into focus, and I think what we're dealing with is essentially a modern view that personal identity equates to identity of consciousness. Who knows, maybe approaching from this angle will provide us with a new way in?

In any case, I think a popular view in this context might be that consciousness is built out of information, and that's what we would be transferring. Unfortunately identity of information doesn't seem to be what we need for personal identity. When we talk about the same

information, we have no problem with it being in several places at once, for example: we don't say that two copies of the same book have the same content but that the identity of the information is different; we say they contain the same information. Speaking of books points to another problem: we can put any information we like down on paper but it doesn't seem to me that I could exist in print form (I may be lacking in energy at times, but I'm more dynamic than that).

So perhaps it's not that the information constitutes the identity; perhaps it simply allows us to reconstruct the identity? I can't be a text, but perhaps my identity could persist in frozen recorded form in a truly gigantic text?

There's a science fiction story in there somewhere about slow transfer; once dematerialised, instead of being beamed to his destination, the prophet is recorded in a huge set of books, carried by donkey across the deserts of innumerable planets, painstakingly transcribed in scriptoria, and finally reconstituted by the faithful far away in time for the end of the world. As a side observation, most transfer proposals these days speak of uploading your consciousness to a computer; that implies that whatever the essential properties of you are, they survive digitisation. Without being a Luddite, I think that if my survival depended on there being no difference between a digitised version of me and the messy analogue blob typing these words, I would get pretty damn picky about lossy codecs and the like.

Getting back to the point, then perhaps the required identity is like the identity of a game? We could record the positions half-way through a chess game and reproduce them on any chessboard. Although in a sense that does allow us to reconstitute the same game in another place or time, there's an important sense in which it wouldn't really be the same game unless it was the same players resuming after an interval. In the same way we might be able to record my data and produce any number of identical twins, but I'd be inclined to say none of them would in fact be me unless the same… what?

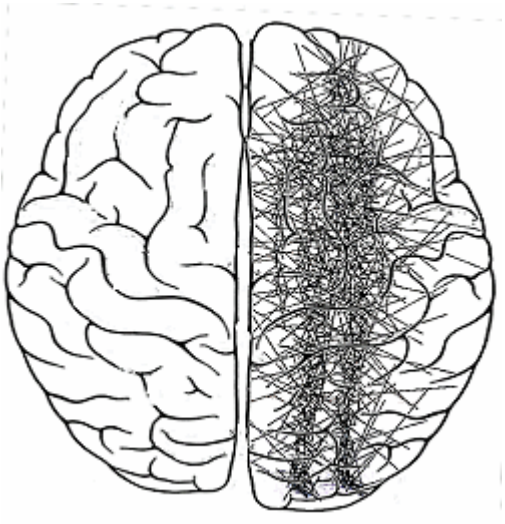
There's a clear danger here of circularity if we say that the chess game is the same only if the same people are involved. That works for chess, but it will hardly help us to say that my identity is preserved if the same person is making the decisions before and after. But we might scrape past the difficulty if we say that the key thing is that the same plans and intentions are resumed. It's the same game if the same strategies and inclinations are resumed, and in a similar way it's the same person if the same attitudes and intentions are reconstituted.

That sounds alright at first, though it raises further questions, but it takes us back towards the quagmire of intentionality, and moreover it faces the same problem we had earlier with information. It's even clearer in the case of a game of chess that someone else could have the same plans and intentions without being me; why should someone built to the same design as me, no matter how exquisitely faithful in the minutest detail, be me?

I confess that to me the whole idea of a transfer seems to hark back to old-fashioned dualism. I think most transferists would consider themselves materialists, but it does look as if what we're transferring is an ill-defined but important entity distinct from the simple material substance of our brain. But I'm not an abstraction or a set of information, I'm a brute physical entity. It is, I think, a failure to recognise that the world does not consist of theory, and that a theory of redness does not contain any actual red, which gives rise to puzzlement over qualia; and a similar failure leads us to think that I myself, in all my inexplicably specific physical haecceity, can be reduced to an ethereal set of data.

Connectome

9 April 2012



Are connectomes the future? Although the derivation of the word "connectome" makes no sense – as I understand it the "-ome" bit is copied from "genome", which in turn was copied from "chromosome", losing a crucial 's' in the process * - it was coined simultaneously but separately by Olaf Sporns and Patric Hagmann, so it is clearly a word whose time to emerge has come.

It means a functionally coherent set of neural connections, or a map of the same. This may be the entire set of connections in a brain or a nervous system, but it may also be a smaller set which link and work together. There is quite a lot going on in this respect: the Human Connectome Project is preparing to move into its second, data-gathering phase; there's also the (more modest or perhaps more realistic) Mouse Connectome Project. One complete connectome, that for the worm Caenorhabditis elegans, already exists (in fact I think it existed before the word "connectome") and is often mentioned. The Open Connectome Project has a great deal of information about this and much besides.

The idea of the connectome was given a new twist by Sebastian Seung in his TED talk "I Am My Connectome", and he has now published a book called (guess what) Connectome. In that he gently and thoughtfully backs away a bit from the unqualified claim that personal identity is situated in the connectome of the whole brain. It's a useful book which falls into three parts: a lucid exposition of the neural structure of the brain, some discussion and proposals on connectomic investigation; and some more fanciful speculation, examined seriously but without losing touch with common sense. Seung touches briefly on the spat between Henry Markram and Dharmendra Modha: Markram's Blue Brain project, you may recall, aims to simulate an entire brain, and he was infuriated by Modha's claim to have simulated a cat brain on the basis of a far less detailed approach (Markram's project seeks to model the complex behaviour of real neurons: Modha's treated them as standard nodes). Seung is quite supportive of these simulations, but I thought his discussion of the very large difficulties involved and the simplifications inherent even in Markram's scrupulous approach was implicitly devastating.

What should we make of all this connectome stuff? In practical terms the emergence of the term "connectome" adds nothing much to our conceptual armoury: we could and did talk about neural networks anyway. It's more that it represents a new surge of confidence that neurological approaches can shoulder aside the psychologists, the programmers, and the philosophers and finally get the study of the human mind moving forward on a scientific basis. To a large extent this confidence springs from technical advances which mean it has finally begun to seem reasonable to talk about drawing up a detailed wiring diagram of sets of neurons.

Curiously though, the term also betrays an unexpected lack of confidence. The deliberate choice of a word which resembles one from genetics and recalls the Human Genome Project clearly indicates an envy of that successful field and a desire to emulate it. This is not the way thrusting, successful fields of study behave; the connectonauts seem to be embarking on their

explorations without having shed that slightly resentful feeling of being the junior cousin. Perhaps it's just that like most of us they are slightly frightened of Richard Dawkins. However, it could also be a well-founded sense that they are getting into something which is likely to turn out complicated in ways that no-one could have foreseen.

One potential source of difficulty lies in the fact that looking for connectomes tends to imply a commitment to modularity. The modularity (or otherwise) of mind has been extensively discussed by philosophers and psychologists, and neurologists have come up with pretty strong evidence that localisation of many functions is a salient feature of the brain: but there is a risk that the modules devised by evolution don't match the ones we expect to find, and hence are difficult to recognise or interpret; and worse, it's quite possible that important functions are not modularised at all, but carried out by heterogeneous and variable sets of neurons distributed over a wide area. If so, looking for coherent connectomes might be a bad way of approaching the brain.

In this respect we may be prey to misconceiving the brain through thinking of it as though it were an artefact. Human-designed machines need to have an intelligible structure so that they can be constructed and repaired easily; and for complex systems modularisation is best practice. A complex machine is put together out of replaceable sub-systems that perform discrete tasks; good code is structured to maximise reusability and intelligibility. But Nature doesn't have to work like that: evolution might find tangled systems that work fine and actually generate lower overheads.

That might be so, but when we look at animal biology the modularisation is actually pretty striking: the internal organs of a human being, say, are structured in a way that bears a definite resemblance to the components of a machine. Evolution never had to take account of the possibility of replacement parts, but (immune system aside) in fact our internal organisation facilitates transplant surgery much more than it need have done.

Why is that? I'd suggest that there is a secondary principle of evolution at work. When evolution is (so to speak) devising a creature for a new ecological niche, it doesn't actually start from scratch: it modifies one of the organisms already to hand. Just as a designer finds it easier to build a new machine out of existing parts, a well-modularised creature is more likely to give rise to variant descendants that work in new roles. So besides fitness to survive, we have fitness to give rise to new descendant species; and modularisation enhances that second-order kind of fitness. Lots of weird creatures that worked well back in the Cambrian did not lend themselves easily to redesign, and hence have no living descendant species, whereas some creature with a backbone, four limbs with five digits each and a tail, proved to be a fertile source of useable variation: leave out some digits, a pair of limbs or the tail; put big legs on the back and small ones on the front, and straightaway you've got a viable new modus operandi. In the same way a creature that bolted on an appendix to its gut might be more ready to produce descendants without the appendix function than one which had reconditioned the function of its whole system (I'm getting well out of my depth here). In short, maybe there is an evolutionary tendency to modularisation after all, so it is reasonable to look for connectomes. As a further argument, we may note that it would seem to make sense in general for neurons that interact a lot to be close together, forming natural connectomes, though given the promiscuous connectivity of the brain some caution about that may be appropriate.

Anyway, what we care about here is consciousness, so the question for us must be: is there a consciousness connectome? In one sense, of course, there must be (and here we stub our toe

on another potential danger of the connectome approach): if we just go on listing all the neurons that play a part in consciousness we will at some point have a full set. But that might end up being the entire brain: what we want to know is whether there is a coherent self-contained module or set of modules supporting consciousness. Things we might be looking out for would surely include a Global Workspace connectome, and I think perhaps a Higher Order Thought Connectome: either might be relatively clearly identifiable on the basis of their pattern of connections.

I don't think we're in any position to say yet, but as a speculation I would guess that in fact there is a set of connectomes that have to act together to support a set of interlocking functions making up consciousness: sub-conscious thought, awareness/input, conscious reflection, emotional tone, and so on. I'm not suggesting by any means that that is the actual list; rather I think it is likely that connectome research might cause us to rethink our categories just as research is already causing us to stop thinking that memory (as we had always supposed) is a single function.

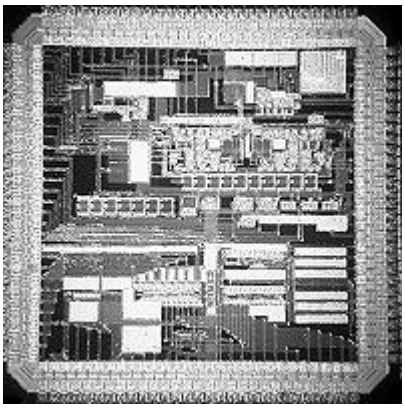
There is already some sign that connectomes might carve up the brain in ways that don't match our existing ways of thinking about it: Martijn van den Heuvel and Olaf Sporns have published a paper which seems to show that there are twelve sites of special interest where interconnections are especially dense: they call this a "rich club", but I think the functional implications of these twelve special zones remain tantalisingly obscure for the moment.

In the end my guess is that by about 2040 we shall look back on the connectome as a paradigm that turned out to be inadequate to the full complexity of the brain, but one which inspired research essential to a much-improved understanding.

**I do realise BTW that words are under no obligation to mean what the Latin or Greek they were derived from suggests – "chromosome" would mean "colour body" which is a trifle opaque to say the least*.

Brain On A Chip

15 April 2012



Following on from preceding discussion, Doru kindly provided this very interesting link to information about a new chip designed at MIT which is designed to mimic the function of real neurons.

I hadn't realised how much was going on, but it seems MIT is by no means alone in wanting to create such a chip. In the previous post I mentioned Dharmendra Modha's somewhat controversial simulations of mammal brains: under his project leadership IBM, with DARPA participation, is now also working on a chip that simulates neuronal interaction. But while MIT and IBM slug it out those pesky Europeans had already produced a neural chip as part of the FACETS project back in 2009. Or had they? FACETS is now closed and its work continues within the BrainScaleS project working closely with Henry Markram's Blue Brain project at EPFL, in which IBM, unless I'm getting confused by now, is also involved. Stanford, and no doubt others I've missed, are involved in the same kind of research.

So it seems that a lot of people think a neuron-simulating chip is a promising line to follow; if I were cynical I would also glean from the publicity that producing one that actually does useful stuff is not as easy as producing a design or a prototype; nevertheless it seems clear that this is an idea with legs.

What are these chips actually meant to do? There is a spectrum here from the pure simulation of what real brains really do to a loose importation of a functional idea which might be useful in computation regardless of biological realism. One obstacle for chip designers is that not all neurons are the same. If you are at the realist end of the spectrum, this is a serious issue but not necessarily an insoluble one. If we had to simulate the specific details of every single neuron in a brain the task would become insanely large: but it is probable that neurons are to some degree standardised. Categorising them is, so far as I know, a task which has not been completed for any complex brain: for Caenorhabditis elegans, the only organism whose connectome is fully known, it turned out that the number of categories was only slightly lower than the number of neurons, once allowance was made for bilateral symmetry; but that probably just reflects the very small number of neurons possessed by Caenorhabditis (about 300) and it is highly likely that in a human brain the ratio would be much more favourable. We might not have to simulate more than a few hundred different kinds of standard neuron to get a pretty good working approximation of the real thing.

But of course we don't necessarily care that much about biological realism. Simulating all the different types of neurons might be a task like simulating real feathers, with the minute intricate barbicel latching structures - still unreplicated by human technology so far as I know - which make them such sophisticated air controllers, whereas to achieve flight it turns out we don't need to consider any structure below the level of wing. It may well be that one kind of simulated neuron will be more than enough for many revolutionary projects, and perhaps even for some form of consciousness.

It's very interesting to see that the MIT chip is described as working in a non-digital, analog way (Does anyone now remember the era when no-one knew whether digital or analog computers

were going to be the wave of the future?). Stanford's Neurogrid project is also said to use analog methods, while BrainScaleS speaks of non-Von Neumann approaches, which could refer to localised data storage or to parallelism but often just means 'unconventional'. This all sounds like a tacit concession to those who have argued that the human mind was in some important respects non-computational: Penrose for mathematical insight, Searle for subjective experience, to name but two. My guess is that Penrose would be open-minded about the capacities of a non-computational neuron chip, but that Searle would probably say it was still the wrong kind of stuff to support consciousness.

In one respect the emergence of chips that mimic neurons is highly encouraging: it represents a nearly-complete bridge between neurology at one end and AI at the other. In both fields people have spoken of 'connectionism' in slightly different senses, but now there is a real prospect of the two converging. This is remarkable - I can't think of another case where two different fields have tunnelled towards each other and met so neatly - and in its way seems to be a significant step towards the reunification of the physical and the mental. But let's wait and see if the chips live up to the promise.

Tallis vs Eagleman

5 May 2012

The Observer has a discussion between Tallis and Eagleman, Eagleman representing neural reductionism and Tallis speaking for a more traditional view of mind and brain.

Although it's worth reading, it turns out a slightly inconclusive encounter. Perhaps on this occasion you'd give Tallis a points victory because he does seem to be looking for a fight, whereas Eagleman is in rather cautious form. They circle each other but never quite identify a proposition which sums up their disagreement clearly enough to get things going.

What seems to emerge is a kind of agreement that mental activity needs to be addressed on more than one level of explanation, with the two antagonists merely giving a different balance of emphasis. This certainly understates the real disagreement between the two.

I think it probably is the case that nearly everyone grants the need for more than one level of explanation. There are those who would say the correct top level is the cosmos itself and that individual consciousness expresses a universal entity. Not quite as high-level as that we surely need to address consciousness on the level of its explicit and social content; we could call this the 'home' level because it is sort of where we live, where we actually experience the world. Most would agree that there are levels of unconscious operation that are also a necessary part of the picture; not many people would say that the structure of the brain and its component neurons tell us nothing; and a majority nowadays would agree that there is ultimately a story at the classical molecular level which, though vastly complex, cannot be ignored. Some say even this is not enough and that consciousness cannot be understood without giving quantum mechanics, or some as-yet-unknown lower level theory, a crucial role.

Only a very hard-line reductionist would say we only need one of these levels: it's generally accepted that there are interesting things to be said on several of them which simply cannot be addressed at other levels. What mainly emerges here is Tallis' defence of the 'home' level against Eagleman's contention that we pay it too much attention and that for many purposes, including our treatment of crime and punishment, we should dethrone it. Intuitively, the motives for Tallis' incredulity are pretty clear: wouldn't it be weird if we had developed the apparatus of thought and consciousness and yet it had no important impact on our behaviour? Don't we just know that discussion and conscious thought ultimately shape what we do, even if our behaviour is sometimes nudged in different directions by factors we're not aware of?

Yet there is something deeply unsatisfactory about the whole idea of different levels of explanation, isn't there? How can one reality require half-a-dozen different accounts? It seems a distressingly messy and arbitrary kind of way for the world to be set up, and certainly we greet any successful reduction of higher level entities to lower level ones as a valuable explanatory achievement. So it's not hard to sympathise with Eagleman's desire to emphasise the role of levels below consciousness either.

Generally speaking it seems that the lower the level of our explanation the better, as though ultimate reality resides at the lowest micro level we can get to. We always celebrate reductions, not elaborations. Yet there have been some rebellious attempts to push things the other way through ideas such as emergence and embodiment, which claim the whole can be more important than the parts. I notice myself that things seem to come most clearly into focus at or slightly below the home level: if we go far above or below that we start to get into regions where we have to deal in probabilities or slightly fuzzy concepts. Most notably there don't seem to be identities at other levels in quite the sharp way there are on the home level. Even molecules are interchangeable: they tell us that it's almost certain we're breathing at least one atom from the oxygen previously breathed by Julius Caesar, but how could you possibly tell? You can't label an atom. Another one of the same kind is distinguishable only by where it is. When we go further down even spatial positions start to get a bit smudged. Equally if we start going up the chain we can only draw slightly fuzzy conclusions about what my family, or the society I live in, thinks or does. This might be a reason to think that real reality is around the home level - or it might a reason to think that the whole business of levels simply flows from my restricted viewpoint and limited understanding.

Perhaps, if my brain were capable of holding it, there is a view on which all the levels could come together. After all, I go on thinking about temperature even though I know it is only molecular motion: perhaps in the end we'll find a way of thinking about the different aspects of mental activity which brings them together without eliminating anything. Perhaps then it might become clear that Eagleman and Tallis don't really disagree at all. I wouldn't put any money on that, though.

The Unity of Consciousness

21 May 2012

The unity of the soul is an ancient doctrine from which we have inherited a strong belief in the unity of consciousness. In certain lights this assumption of unity seems unquestionable, but it has actually been a continual problem; it could almost be argued that the history of understanding the mind has been a history of giving up on unity.

Like other persuasive doctrines that have turned out to be problematic in the long run, we can trace this one back to Aristotle, but it is tied in to a widely-held set of scholastic/ancient ideas about metaphysics. I believe the argument runs more or less like this: the soul is not physical, therefore it lacks extension (which is a physical property); if it lacks extension it necessarily lacks parts, and if it lacks parts it must be single and unified. The soul is a substance, in the old philosophical sense of something incapable of being analysed or broken down. Substances in this sense used to be considered necessary building blocks of reality, required in order to have a secure ontological foothold. Otherwise the process of analysis would be bottomless and unending, and nothing would ever be completely clarified, which would be intolerable (although I notice contemporary physics seems to tolerate a position not altogether unlike this). Readers may well by now feel parts of their own souls waving urgent hands to attract attention to a host of salient objections, but let's avoid getting bogged down in this treacherous territory and move on a bit.

Descartes, say what you will about his dualism, effected a radical change for the better when he restricted the interventions of the soul to the pineal gland: on his view it did its stuff there and the rest of the body worked like a machine, according to the same physical laws as any inanimate stuff. Until then it had been largely assumed that the soul directly activated the body without needing any kind of transmission mechanism. Now I say 'until then', but the remarkable fact is that people went on thinking that way for a long time afterwards. As late as 1850, Helmoltz's measurement of the speed of nerve impulses was resisted by some on the grounds that the vital impulse must act throughout the body simultaneously. When your arm moved, it was because you wanted it to, and it, as part of you, wanted to too. I believe there was a school of thought that held out for a middling point of view, accepting that in principle the brain controlled the body by nerve impulses, but confidently expecting that they would be too blindingly fast to ever be measured. This, of course, proved to be quite wrong, but the nineteenth-century debate is in some ways quite reminiscent of the more recent discussion of Libet. Muller and others thought Helmholtz must be wrong because he was introducing a delay between will and act; people suppose Libet must be wrong because he introduces a delay between deciding to act and awareness. All such delays are intolerable if we insist on the absolute unity of the conscious mind because you can't have a delay between a thing and itself.

Another prominent example of the problems flowing from unity is the vexed issue of the binding problem. Given that sensory inputs come in by different pathways at different speeds and get processed in different ways in different parts of the brain, how is it we end up with a smoothly integrated picture of reality which assigns the right qualities to the right objects and unrolls steadily in real time without jumps, pauses, or lipsynch errors? There are various ways, more or less satisfactory or problematic, in which the brain might ensure everything is properly put

together when it arrives in consciousness, but if we're not assuming consciousness is a single united destination, the problem wouldn't arise in the first place.

Perhaps, though, the binding problem gives us a clue about why unity seems so undeniable - because the contents of consciousness look united. Isn't that it?

Well, sort of, but when I sit down and conscientiously introspect, I don't really detect a lot of unity. At the moment I have strings of explicit words running through my mind a moment before I type them: I moment ago as I sat in uffish thought, I had thoughts about the same subject which were wordless. Half an hour ago I wasn't thinking about anything at all, though I was certainly conscious, and a bit before that I was largely absorbed by experiencing the taste of scrambled egg. An hour before that I was dreaming and some time before that in a blank state of which I can't say for sure whether I remember it or not.

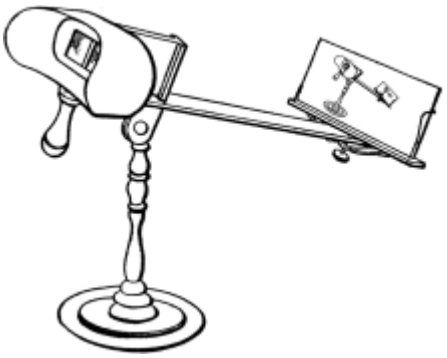
It's worse than that, because at all these times there were also things in the penumbra of my mind which I was aware, or perhaps only pre-aware or potentially aware of. Hume famously said that when he looked into his mind he found only a bundle of sensations; but how simple it would be if the sensations were really always bundled; if they were always of the same broad kind; and if they were all merely sensations, instead of including bits of broken intentionality, fragments of half-or potentially meaningful intimations, things that might be the phenomenally detectable end of affordances, incipient recognitions and implicatures and an exquisitely ineffable and shadily located intimation that there may soon be the emergence of what we can call a gut feeling finely balanced on the cusp between the affective and the merely digestive. A bundle? Really a heap, or even a cloud, would be far more orderly and unified than my subjective experience.

What does bind things together is a kind of bird's nest framework of memory linking now to then, and then to some other experience, and so on; but this is not all that useful. For one thing these linkages are loose - what they provide is the kind of tangled and ad hoc unity the cables behind my stereo have achieved unbidden - and they are fallible, not part of the essence of consciousness. I see nothing absurd about the idea of my having moments of consciousness which are neither remembered nor involve remembering.

Yet even so I find it intuitively impossible to abandon the idea of some unity somewhere, even if I can't quite put my finger on it at the moment. Things would be so much easier if we didn't exist, but there we are.

HOT Death

1 June 2012



You may have seen the very interesting review that Micha kindly mentioned recently. I plan to discuss that next week, but before talking about Higher Order Theories (HOTs) it seemed best to set the context by talking about Ned Block's paper which seeks to demolish them.

Higher Order theories come in many flavours, but the basic proposition is that a mental event, a thought or a feeling, is conscious if there is another thought about it, that original mental event. The basic intuition is that you can be aware of something unconsciously, but when you're aware of your awareness it's conscious.

Not everyone likes this perspective (Roger Penrose caustically pointed out that pointing a video camera at itself doesn't make it conscious) and some would say either that it only explains certain varieties of consciousness or that it only explains certain aspects of consciousness. Nevertheless, HOTs have had a long run as respectable contenders, representing one of the major areas where we might choose to look for The Answer.

Block's paper last year was unusual in seeking to offer something like a knock-down destruction of the case, or perhaps it would be more accurate to sought he meant to pursue the HOTists until their last hiding place had been flushed out. His targets were not the modest theorists who claim that HOTs may explain some varieties of aspects of consciousness, but the ones he characterised as 'ambitious': in particular those who claim HOTs could explain the 'what-it- is-likeness' (let's call it WIIL) of conscious experience.

The starting point is the uncomfortable fact that we can have thoughts about being in conscious states we're not in fact in. We can think we're seeing red when in fact we're seeing green, or not seeing anything at all. You might well feel that this in itself is something of a blow for HOTs, and your first reaction might be that they should give up any claim of a conscious state where there is no state to work with. That sounds sensible but the retreat is not so easy as it seems if we want to retain WIIL for dreams and illusions, as we surely do. In any case, Block's targets take the opposite path: sure there can be WIIL in those cases, they reply.

Now Block springs the trap. So you're saying, he observes, that an episode is conscious if it is the object of a simultaneous higher order thought? And that thought is a sufficient condition for a conscious episode. Yet it's also a necessary condition that that episode is the object of a higher order thought. Yet in this case we have only the one, higher-order thought to work with (and we can assume it ain't self-referential). So there is no conscious episode. We've got necessary and sufficient conditions which are not compatible – what madness is this?

There is a way out which those unused to philosophical discussion may find a little odd: this consists of saying that in these cases, where we have a second-order thought about an experience we're not actually having, there is an object of the second order thought after all: it's simply one that doesn't exist, that's all. We must remember here that the objects of thought are slippery customers; we often think about things that don't exist (Pokemon, the sixth wife of Henry VII, the house I would have built if I had won the lottery, square circles).

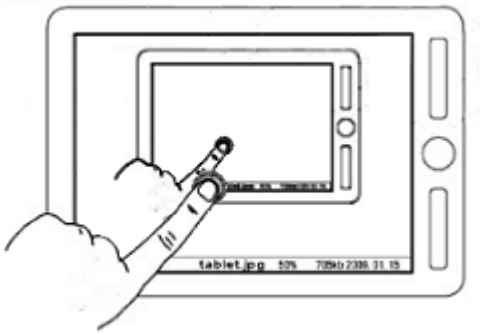
But, says Block, if you take that route, where's your WIIL (or maybe in this case it should be WIIAL –What It Isn't Actually Like)? It's now fake WIIL – and you can't rest content with that.

I've omitted some important technicalities and details in the foregoing, but I hope the gist comes through: where does it leave us? It seems an effective argument to me, and I would add that for those who are seeking the essence of WIIL, it seems intuitively unlikely to me that it could ever have resided in thoughts about thoughts: that brings it all back into the head, whereas what it is like ought to out there with 'it'. Those who never believed in WIIL will not, of course, be troubled by any of this.

So following Block's demolition, the advocates of HOT admitted their error, thanked him for clarifying, and issued a full retraction. No, of course they didn't, as we shall see.

Still HOT

10 June 2012



So now let's look at the review by Richard Brown of Rocco J. Gennaro's new book, which sets out his HOT. I should disclose that I have not read Gennaro's book, so I'm merely reviewing a review of a theory. You could call it a higher order review.

Gennaro's own concerns feature concepts strongly: his theory includes conceptualism, the belief that concepts provide key structure for our conscious experience. This causes him some difficulty over animals and infants, as he needs them to have sufficiently advanced concepts to form the required higher order awareness. He needs at least some account of how the required concepts are acquired (or how they came to be innate), and he needs to avoid getting into the bootstrap bind where you need to have sophisticated conceptual structures in order to pick up the concepts you need in order to have sophisticated conceptual structures.

My personal bias is that conceptualism tends to bring unnecessary complications, and in this particular case I'm not sure why we need to make the weather so heavy (though no doubt Gennaro has his reasons). All we need is the awareness of an awareness: the higher order awareness does not need to look through to the objects in the external world. The conceptual apparatus required surely ought therefore to be modest, and I wouldn't fight very hard against the idea that it could be built in or be the result of a very early internal rewiring exercise in the infant mammalian brain.

The main point of interest for us is the difference which Brown sets out between his own view and Gennaro's. This hinges on a fundamental point: are we talking about higher order awareness of a mental state (Gennaro), or of our being in that mental state (Brown)? At this point I'd say I find Brown's view, which he characterises as the non-relational view, more appealing. It seems important to specify that the awareness involved must be ours, after all.

However, both sides claim that their view is best able to deal with the kind of objections raised by Block and discussed last time, arising from cases where we have the higher order awareness but actually through some error the lower-order awareness which it targets is not actually there. Gennaro, it seems, wants to disqualify these states altogether: where the first-order state is not there, there's no consciousness. If you're not seeing something red, you can't have the subjective consciousness of redness. This is neat in its way, but seems arbitrary (How come subjective experience, of all things, comes to have a special kind of immunity from error?), and surely we want to retain the possibility of a subjective experience not based in reality? I can see that some might argue that dreams lack real subjective experience, but are we prepared to say the same of illusions and mirages? That seems a high price to be paying.

Brown's escape route is quite different: by adopting the non-relational view he can cut free from the first-order state altogether. Who cares whether it's there or not? It's just about the right kind of second-order state, that's all. This may seem a little weird, but after all the orthodox view of qualia, our subjective experiences, is that they are largely decoupled from ordinary causal reality. On Brown's view, if we see green, we can have the experience of red without invalidating the experience. We'll still behave as though we were having the experience of green. Block

might denounce our qualia as fake, but meh, if that's what you mean by fake, all qualia were always fake, so who cares?

The logic of that seems faultless, but I couldn't help feeling that my sympathies were swinging back to Gennaro somewhat. Brown mentions a second problem, the problem of the rock: why can't we make a rock conscious by having the right kind of second-order mental state about it? For Brown's decoupled view, there's no problem because we're not even talking about the first order awareness - rock, schmock, it's simply irrelevant. Gennaro does have to face the issue and it seems he seeks to do it simply by disqualifying rocks with an additional rule or specification. Brown considers this another unattractively arbitrary lash-up, and perhaps it is, but in another light it seems far closer to common sense (though of course the realm of common sense is some way off by now in any event).

For myself the net effect of the discussion is to make me feel more strongly than before that if we are to have a HOT (and I'm not absolutely wedded to that in itself), we'd do much better to stick with what Block calls the unambitious variety, the kind that doesn't seek to explain subjective experience or exorcise those deadly sirens we call qualia.

Subliminal Revolution

24 June 2012

The subtitle of Leonard Mlodinow's book *Subliminal* makes bold claims: *The Revolution of the New Unconscious and what it Teaches us about Ourselves*.

Mlodinow is a talented fellow: I first became aware of him as Stephen Hawking's co-author on The Grand Design (I blamed him then for the terrible jokes in that book, but the evidence of Subliminal, which is amiable but wince-free throughout, I think Hawking was probably to blame for them after all). Being Hawking's colleague is probably the nearest the modern world can offer to being God's assistant, but in addition Mlodinow has done impressive original work in physics and written successful screenplays.

The book is a wide-ranging compilation of a lot of interesting stuff. In the early stages of the book, it seems Mlodinow is basing his claims on contemporary technology and fMRI in particular: he tells us it is transforming our knowledge. But in fact not much of the research he reports is dependent on scanning. It feels as if the book might have changed direction in the writing, as Mlodinow found that most of the stuff he wanted to include actually didn't involve advanced technology after all, but retained in the text the laudatory stuff about fMRI which it no longer really justifies.

How come the scanners don't feature more strongly? One possible reason is sort of indicated when Mlodinow talks about how experimental subjects were shown to rate wine more highly when told it was expensive. Mlodinow wants to say that the tasters did not merely give the 'expensive' wine better ratings, but actually enjoyed it more: so he tells us that fMRI scans showed activity in the orbitofrontal cortex, 'a region that has been associated with the experience of pleasure.' Has been associated (not necessarily by me) associated with (not necessarily controlling or unambiguously diagnostic of). If the trumpet sounds as uncertainly as that, we must ask whether whether we're really being told anything of value. Of course we know why Mlodinow is so hesitant. First, nobody has a really clear idea of what the orbitofrontal cortex does; it seems to be involved in addiction and motivation - but for Mlodinow's purposes we need to be talking about the qualia-laden appreciation of fine bouquets and the like, which may well be an unrelated matter. Second, fMRI is a fuzzy and ambivalent tool and the wider implications of the data it produces are always debatable. Third, this business of real pleasure is a philosophical swamp: put aside all the Hard Problem issues: were the subjects experiencing real pleasure, or did they just think they were experiencing pleasure, or were they just thinking about pleasure? Was the pleasure straightforwardly gustatory, or did it come from thinking what smart guys they were and wishing their friends could see them now? These are not mere quibbles; the latter case for example, would be much less interesting than the radical and somewhat implausible claim that beliefs about price really change the experience.

That points to a general difficulty: books of this kind often give us stuff that is interesting, new, and well-founded; but the stuff that is well-founded isn't new and the stuff that is interesting is debatable and looks over-interpreted. I wouldn't say Mlodinow escapes this pitfall entirely. He

tips his hat generously to Freud, which is nice, but that's surely the Old Unconscious. The wine experiment - how many eyebrows would that have raised around the table at Plato's symposium? Perhaps not many. Mlodinow tells us yet again the story of how Nixon lost out to Kennedy in 1960; people who could see him on TV were less inclined to think he had won the debate than those who merely heard him on the radio. Well, we've known that people are influenced by candidates' appearance at least since Pericles took to appearing in a helmet which both reminded the electorate of his generalship and concealed the weird shape of his head. Do we even know that people were unconscious of being influenced by Nixon's appearance? It seems quite possible that some of them drew the entirely conscious conclusion that he looked too rough and too shifty to be credible (a verdict which some would argue was borne out by later history, incidentally). On the other hand, Mlodinow reports research showing that people named 'Brown' are significantly more likely to marry other people called 'Brown' than statistical chance would warrant. Is that true? Or is there some quirk here - perhaps there are 'Brownsvilles' where by chance or history a concentration of Brown families mean you're more likely to meet people of that name than random population matching would suggest? I don't know, but I'm left in doubt, and as a human being myself I need something pretty strong to convince me to give up my strong intuitive understanding that surnames are not generally relevant to my species' mating decisions.

The point that electors may have assessed Nixon's appearance emotionally but consciously leads us to another difficulty: quite a bit of the research Mlodinow recounts doesn't really bear on his thesis about the unconscious. He recounts the experiment, by now fairly well-known, in which an experimenter asked a stranger for directions: accomplices interrupted the conversation by carrying a door between the two, behind which the experimenter was switched for someone else: subjects often resumed the conversation without noticing the change in their conversational partner (the book here sort of undercuts the experiment by including pictures in which it is clear that the two experimenters were not that dissimilar in looks, and, if I may be rude, also of a rather unstriking generic appearance, too).

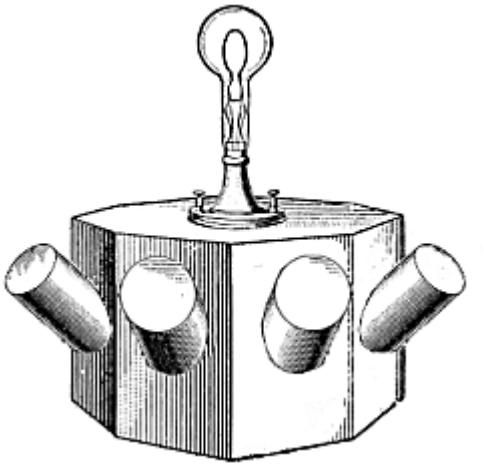
The experiment is interesting, but how does it show that the unconscious is more important than we thought? Is there any suggestion that the difference was recognised unconsciously while being ignored consciously? Well, no: in fact we might think that this is the sort of thing the conscious wouldn't deal with, leaving itself to be warned by unconscious processes, so if anything the hit is against the effectiveness and influence of the unconscious. Simply showing errors in conscious beliefs does not establish a revolution in favour of a new unconscious.

But then Mlodinow never formulates what he means by either the old or the new unconscious. We don't even know whether he thinks the unconscious really amounts to one thing, several different unconsciousnesses, or simply a lot of default non-conscious mechanisms. The word 'consciousness' notoriously covers a number of different entities or processes, but we never get told explicitly which of them Mlodinow believes in or which of them he wants to dethrone. If you want to carry out a revolution against one form of mental activity and in favour of another, you really need to offer a pretty clear of view about what those different forms actually are and what roles they play, don't you? Mlodinow would never try to get away with such vagueness if he were trying to sell us a revolution in physics, so the fact that he seems to think it will do for consciousness suggests an unattractive casualness, to say nothing worse. Perhaps in a way it's evidence in his favour that Mlodinow never seems to have noticed consciously that so much of his material doesn't really bear on his thesis; perhaps his unconscious is subtly offering us a different verdict.

That may be just a little hard: there's a lot of very readable stuff about genuinely interesting research here, but the Revolution of the New Unconscious seems to me to have gone missing.

A New Light On Computation

8 July 2012



A couple of years back I mentioned some ideas put forward by Mark Muhlestein: a further development of his thinking has now been published in Cognitive Computation.

The paper addresses the question of whether a computational process could support consciousness. Muhlestein begins by briefly recounting some thought-experiments proposed by Maudlin and others. Suppose we run a computation which instantiates consciousness: then we run the same process but remove all the unused conditional branches, so that now the course of the process is fixed. In the second case the computer goes through the same set of states (does it, actually?), but we'd be disinclined to think it could be conscious; it's just a replay. In fact, it's hard to see why 'replaying' it makes any difference since the relevant states all exist in the record, so that inert record itself probably ought to be conscious. Worse than that, since we could, given a sufficiently weird and arbitrary decoding, read random patterns in rocks as recording the same states, an indefinite number of copies of that same consciousness would be occurring constantly pretty much everywhere.

That's more or less how the argument runs, anyway, and what's proposed to fix the problem is the idea that for consciousness to occur, the computation must have the property of counterfactual sensitivity. It must, in other words, have had the capacity to go another way. Without that property, no consciousness. The notion has a certain intuitive plausibility, I think: we can accept that in order for me to have the experience of seeing something, the fact that it was a red square and not a blue circle must be relevant; and that consciousness perhaps must be part of a stream which is, in some unobjectionable sense, free-flowing.

Muhlestein proposes a new thought experiment with a truly formidable apparatus. He set up a physical implementation of Conway's Game of Life and uses that in turn to implement a computer on which his processes can be run. Because his implementation of Life uses cellular automata which display and detect states using light, he can now intervene anywhere he likes by simply shining an external light into the process.

If that last paragraph is unintelligible, don't worry too much: all you need to know for the purposes of the argument is that we have a computer set up so that we have a special power to intervene arbitrarily from outside and constrain or alter the process which is running as it progresses. Muhlestein now takes a conscious computational process (in fact he proposes to scan one on a whole-brain basis from his friend Woody - don't try this at home, kids!) and runs it on his set-up; but here's the cunning part: he uses his light controls not to interfere, but to project the same process back onto itself. In effect, he's over-determining the process: it runs normally by itself, but his lights are enforcing a recorded version of the same process at the same time.

Now, the computation is taken to be conscious when running normally. It runs in exactly the same way when the lights are on; it simply loses the counterfactual sensitivity: it could no longer have gone another way. But why would extra lights on the process deprive it of

consciousness? The outputs will be exactly the same, any behaviour of the conscious entity will be the same, and so on. Nothing about the fluidity or coherence of the process changes. Yet if we say it remains conscious, we have to give up the idea that counterfactual sensitivity makes a difference, and then we're back in difficulty.

What do we say to that? Muhlestein ultimately concludes that there is no satisfactory way out short of concluding that in fact the assumptions are wrong and that computational processes are not sufficient for consciousness.

Old Peter (the version of myself that existed a couple of years ago) thought the problem was really about appropriate causality, though I don't think he explained himself very cogently. He would rightly have said that bringing counterfactuals into it is dangerous stuff because they are a metaphysical and logical minefield. For him the question would be: do the successive states of the computation cause each other in appropriate ways or do they not? If they do, we may have a conscious process; if the causal relations are disrupted or indirect, we're merely waving flags or playing with puppets. So in his eyes the absence of counterfactual sensitivity in Muhlestein's example is not decisive and his lights do not remove the consciousness unless they disrupt the underlying causality of the computation: unless they make a difference, in short. Causal over-determination of the process is irrelevant. The problem for Old Peter is in explaining what kinds of causal relations between states have to obtain for the conscious computation to work. His intuition told him that in real processes, including those in the brain, each event causes the next directly, whereas in simulations the causal path is through a model or a script. Unfortunately we could well argue that this kind of indirect causal path is also typical of computation, leading us back to Muhlestein's conclusion by a different route. Myself, I'm no longer completely sure either that it is a matter of indirect causality or that computational processes are necessarily of that kind.

For myself I should still be inclined towards suspicion of counterfactual sensitivity, but I would be more inclined to say that what we're missing is the element of intentionality; before we can complete the analysis of the problem we need to know what it is that endows a brain process with meaning and aboutness. The snag with that strategy is that we don't know.

All in all, I found it interesting to look back a bit. It seems clear to me that over the years this site has been up I have made some distinct progress with consciousness: every year I know a little less about it.

The Stove of Consciousness

23 July 2012

I've been reading A.C. Grayling's biography of Descartes: he advances the novel theory that Descartes was a spy. This is actually a rather shrewd suggestion which makes quite a lot of sense given Descartes' wandering, secretive life. On balance I think he probably wasn't conducting secret espionage missions - it's unlikely we'll ever know for sure, of course - but I think it's certainly an idea any future biographer will have to address.

I was interested, though, to see what Grayling made of the stove. Descartes himself tells us that when held up in Germany by the advance of winter, he spent the day alone in a stove, and that was where his radical rebuilding of his own beliefs began. This famous incident has the sort of place in the history of philosophy that the apple falling on Newton's head has in the history of science: and it has been doubted and queried in a similar way. But Descartes seems pretty clear about it: *"je demeurais tout le jour enfermé seul dans un poêle, où j'avais tout le loisir m'entretenir de mes pensées".*

Some say it must in fact have been a bread-oven or a similarly large affair: Descartes was not a large man and he was particularly averse to cold and disturbance, but it would surely have to have been a commodious stove for him to have been comfortable in there all day. Some say that Bavarian houses of the period had large stoves, and certainly in the baroque palaces of the region one can see vast ornate ones that look as if they might have had room for a diminutive French philosopher. Some commonsensical people say that "un poêle" must simply have meant a stove-heated room; and this is in fact the view which Grayling adopts firmly and without discussion.

Personally I'm inclined to take Descartes' words at face value; but really the question of whether he really sat in a real stove misses the point. Why does Descartes, a rather secretive man, even mention the matter at all? It must be because, true or not, it has metaphorical significance; it gives us additional keys to Descartes' meaning which we ought not to discard out of literal-mindedness. (Grayling, in fairness, is writing history, not philosophy.)

For one thing Descartes' isolation in the stove functions as a sort of thought-experiment. He wants to be able to doubt everything, but it's hard to dismiss the world as a set of illusions when it's battering away at your senses: so suppose we were in a place that was warm, dark, and silent? Second, it recalls Plato's cave metaphor. Plato had his unfortunate exemplar chained in a cave where his only knowledge of the world outside came from flickering shadows on the wall; he wanted to suggest that what we take to be the real world is a similarly poor reflection of a majestic eternal reality. Descartes wants to work up a similar metaphor to a quite different conclusion, ultimately vindicating our senses and the physical world; perhaps this points up his rebellion against ancient authority. Third, in a way congenial to modern thinking and probably not unacceptable to Descartes, the isolation in the stove resembles and evokes the isolation of the brain in the skull.

The stove metaphor has other possible implications, but for us the most interesting thing is perhaps how it embodies and possibly helped to consolidate one of the most persistent metaphors about consciousness, one that has figured strongly in discussion for centuries, remains dominant, yet is really quite unwarranted. This is that consciousness is internal. We routinely talk about "the external world" when discussing mental experience. The external world is what the senses are supposed to tell us about, but sometimes fail to; it is distinct from an internal world where we receive the messages and where things like emotions and intentions have their existence. The impression of consciousness being inside looking out is strongly reinforced by the way the ears and the brain seem to feed straight into the brain: but we know that impression of being located in the head would be the same if human anatomy actually put the brain in the stomach, so long as the eyes and ears remained where they are. In fact our discussions would make just as much sense if we described consciousness as external and the physical world as internal (or consciousness as 'above' and the physical world as 'below' or vice versa)

If we take consciousness to be a neural process there is of course, a sense in which it is certainly in the brain; but only in the sense that my money is in the bank's computer (though I can't get it out with a hammer) or Pride and Prejudice is in the pages of that book over there (and not, after all, in my head). Strictly or properly, stories and totals don't have the property of physical location, and nor, really, does consciousness.

Does it matter if the metaphor is convenient? Well, it may well be that the traditional inside view encourages us to fall into certain errors. It has often been argued (and still is) for example that because we're sometimes wrong about what we're seeing or hearing, we must in fact only ever see an intermediate representation, never the real world itself. I think this is a mistake, but it's one that the internal/external view helps to make plausible. It may well be, in my opinion, that habitually thinking of consciousness as having a simple physical location makes it more difficult for us to understand it properly.

So perhaps we ought to make a concerted effort to stop, but to be honest I think the metaphor is just too deeply rooted. At the end of the day you can take the thinker out of the stove, but you can't take the stove out of the thinker.

Nothing Could be Better...

6 August 2012

We discussed once before that old philosophical puzzler: why is there actually anything at all? Jim Holt's new book Why does the World Exist? is an entertaining but basically serious assault on this fundamental issue.

Near the beginning he has a splendid chapter about nothing (as it were). He's a little hard on the Greeks and Romans, suggesting that to them the very idea of zero was inconceivable. That's surely reading too much into the fact that Roman numerals don't include a symbol for zero (and you know, the Romans themselves didn't even use those formal numerals when they needed to do everyday sums, but another system altogether). But his brief history of nothing from Heidegger ('nothing noths', apparently) through Bergson to Nozick, and his explanation of the problems that arise from confusing nothing and nothingness, is lively in the best kind of way and stimulating. Could nothing even exist? Holt regards true nothingness as tough to conceive of and spends some time on different ways of attempting the feat. I don't think I find it that hard myself to think that the world might be null and empty or a Euclidean point (there's a Greek conception that's pretty close to zero for you).

That is the question that drives this enquiry, though: why all this stuff? Wouldn't it be more natural if there was nothing? Much of the book is formed of conversations with selected luminaries, and the first of these, Adolf Grünbaum, simply denies that there's anything puzzling about the world's existence. All this stuff about creation of a world out of the void is just a hangover from Genesis as far as he's concerned. That seems a little too easy. Holt's second interlocutor, Richard Swinburne, thinks by total contrast that the simplest explanation for the cosmos is, in a word, God, though God himself is inexplicable. That, in a quite different way, seems too easy too.

Holt talks to David Deutsch about a quantum multiverse and Steven Weinberg about Theories of Everything, but both, with refreshing clarity and honesty, deny possessing any ultimate answer to his question. Roger Penrose believes in three separate worlds, but his basically Platonic conception doesn't seem to offer us anything very new, or to me very satisfying.

With John Leslie things start to get interesting, if bizarre: he believes in axiarchism: the universe exists because of the moral requirement that goodness should exist. There seem to be great difficulties with getting ontology from ethics: attentive observers will have noticed that the moral requirement for goodness doesn't seem to have much direct causative effect in the real world.

The person Holt is most impressed by is Derek Parfit, who appears (I haven't read Parfit himself on the subject) to offer not so much a theory as a framework in which all possible universes are theoretically available, but one is actualised by a Selector, a principle which prioritises one. Holt likes this framework and he builds on it a theory of his own. For reasons which were never clear to me, he believes we only need to consider four possible Selectors: simplicity, goodness, fullness/non-arbitrariness, and no Selector at all. Surely there are many more possibilities than that, whatever a Selector is supposed to be (and that's not quite clear either: for Parfit I suspect

it is the kind of intermediate convenience that can be cancelled out of the final solution but Holt seems at times to take it as a metaphysical reality)? Anyway, Holt decides to arbitrate between the Selectors by using them on themselves as meta-Selectors. He thinks only two emerge from this exercise: Simplicity and Fullness. He concludes that if all selectors or no selectors are going to be applied as a result, we end up with a universe of surpassing mediocrity. That seems to be his final view: the world exists like this because it was the most mediocre option the cosmos could come up with. Amusingly he goes on to ask: what could be the reason for my own existence in such a Universe?

Not a convincing conclusion, then, but an intelligent account of the kind that causes the reader to nod in sage agreement or exclaim in frustration by turns.

Turning aside from Holt's conversations, let's see if we can get straight the reasons for puzzlement about why there is anything. Well, isn't it that Occam's Razor tells us that entities are not to be multiplied, and so we expect a minimal number of entities in our world. Why does Occam work? I think it really operates on two levels. Strictly Occam is a metatheoretic principle of parsimony: it's not about reality, it's about which theory we should prefer. Given any set of facts, there is an unlimited number of different theories which will account for them: we need some way of choosing and Occam tells us to pick the simplest account, or more precisely, the one whose picture of the world contains the smallest and least complex set of things. Strict Occam doesn't tell us that this simplest theory is absolutely going to be the true one, and sometimes it eventually turns out that it isn't; but going for simplicity seems the most practical way of picking out one version from the range of possible theories: it may be the only fully comprehensive and consistent principle we can apply. (In practice what we often rely on is not a marginal difference of complexity but a kind of Occamic click: sometimes when a good theory comes along it provides an abrupt and substantial drop in the complexity of our world view: all of a sudden quite lot of different things make sense, and that's really what makes us think the theory must be true.)

But in addition there is second pseudo-Occamic principle lurking below the surface: things don't happen/exist for no reason. Whereas the strict version of Occam is epistemic, this idea is ontological: it doesn't just tell us what to think it tells us what probably exists, and prompts us to think that the simplest theories are not just the most convenient to adopt, but more likely to be true than any others. It is mainly this pseudo-Occamic idea that leads us to be, not happy as kings (as Robert Louis Stevenson suggested) but puzzled that the world is so full of a number of things. Leslie's axiarchism seems almost to invert this principle, claiming that good stuff can indeed spring into existence for no other reason than its intrinsic ethical qualities.

We should notice that the pseudo-Occamic principle mainly tells us that things don't change; so if there were nothing to begin with, nothing could be expected to come of it (a point Holt covers in his discourse on nothing, by the way); but even if there is something, we should expect the universe to be relatively uneventful. So even if it's easy for Grünbaum to shrug off our Occamic propensity to prefer a theory which gives us a cosmic nothing, he also needs to explain why we seem to be in the midst of a world which has a long complex history and is constantly changing.

One possible answer here is some form of the Anthropic Principle: we're in a world with that long complex history because otherwise it couldn't contain us. As we know, the anthropic principle comes in a variety of forms: at one end there are unobjectionable versions which say: the fact that we find ourselves in a world that contains enough detail for human beings to be

part of it is no more surprising than the supposed stroke of luck that we got ourselves born on a planet with an atmosphere: if it had been otherwise, we wouldn't be here to talk about it. At the other extreme are scarily idealist versions of the principle which say that our existence actually reaches back in time and reconfigures the fundamental constants of the universe.

Another reason we might want the anthropic principle is to help us whittle down a multiverse. Multiverses of one kind or another seem to be a popular option for helping to explain the world, but again they come in reasonable and less reasonable forms. Parfit seems to have one variety in which slightly different laws or constants operate in different regions of space; but other conceptions have sets of universes which implement everything that isn't logically contradictory. There are problems with this, not least with the issue of identity across worlds. Multiversians would have it that are universes where I never wrote this post - but there aren't, because I did: those people who failed to write it are people exactly like me, not me: and it follows that all possibilities are not realised after all, and cannot be, no matter how much the worlds proliferate. Personally I suspect that since the alternative universes make no difference to our world, merely providing an explanatory convenience which they could do even if they didn't in fact exist, they might as well not exist.

Anyway, what has all this got to do with consciousness? Puzzlement over the existence of the world is partly, I submit, puzzlement over why there are such specific and apparently arbitrary details to it. Why anything, yes, but even if something,why on earth all this? That is strongly related to the questions why me? and What on earth am I? There is a special intractability to questions of this kind: we don't want the answer to be purely logical because then we would get eternal archetypes, and we're not that; but we don't want something random, arbitrary, or notional either because after all we are real. Theories are about generalities, but we're asking for a theory of the particular. It is haecceity - thisness - that makes us and our qualia so special, but haecceity seems to require an unprecedented kind of explanation that doesn't exist.

Next time I'll attempt to give it.

...but Something Is Necessary

19 August 2012

Last time I suggested that we might approach the Hard Problem of qualia by first solving the impossible problem of why the world exists at all (what the hell, eh?). How would that work?

Qualia, of course, are the redness of red, the indescribable smelliness of the smell of fish, and so on; the subjective, phenomenal, inexpressible qualities of experience, the bit that the scientific account always leaves out. They are often described as the 'what it is like' of an experience, and have been memorably characterised as what Mary, who knows everything about colour, learns when she actually sees it for the first time.

My case is that a large part, perhaps all, of the strange ineffability of qualia arises because what we're doing is mismatching theory and actuality. It should not really be a surprise that the theory of red coloration does not itself deliver the actual experience of redness, but there is some mysterious element in actual real-life experience that puzzles us. I suggest the mysterious extra is in fact haecceity, or thisness; the oddly arbitrary specificity of real life, which sits oddly with the abstract generalities of a theoretical description. So it would help to know why the actual world is so arbitrary and specific: why it isn't a featureless void, or a geometric point, or a collection of eternal Platonic archetypes. If we knew that we might know something about qualia; and also, I think about ourselves, since we too are arbitrary and specific, not abstract functions or sets of information, but real one-off items.

So can we therefore answer Jim Holt's question for him? Some caution is certainly in order. Speculative metaphysics is like hard drink; a little now and then is great, but you need to know when to stop or you may find your credibility, if not your coherence, diminishing. But I think we can sketch out a tentative view which will clarify a few points and indicate some promising lines of inquiry which may well be rather helpful.

Let's step back and look at the overall cosmic problem more empirically: the world does in fact exist and does seem to be rich and complex. What kind of overall chronology would make sense for a world like that? It could be one that starts, putters along for a finite time, and then stops. It could be one that stretches back indefinitely into the past but eventually stops at some point in the future, or one that started at a definite point in the past but goes on indefinitely. It could be indefinitely prolonged at both start and end. Or it could be one that goes round in a permanent loop.

The thing is, in different ways all these options seem to give us a universe which is unmotivated. If there is a final state in which the universe stops, why not go to that state in the first place – why spend time getting there? If the universe goes in a circle and ultimately reaches the state in

which it started, why bother leaving the initial state? Our current view paints a picture of a Universe wound up by a cataclysmic Bang and then steadily running down through expansion and increasing entropy, to nothing, or as near nothing as makes no great difference: but it seems odd to start the world with a flagrant contradiction of the principle of decline that afterwards governs its development. The only world that makes sense in these admittedly vague terms is one that is going somewhere, but somewhere it will never finally reach: the only one that does that, I think, is one that starts and then goes on, not merely expanding, but transcending its earlier states and rising to higher levels of complexity indefinitely.

That doesn't quite tell us why there is anything at all. What caused this indefinite existential transcendence in the first place? Some kind of ontic horror vacui? An inherent cosmic desire for there to be more stuff? One of the things I noticed in reading Jim Holt's book was that there are one or two gaps in our conceptual toolkit when we embark on this quest. One of those is that we are looking for a causal explanation but we don't really have any clear idea of what causation is at a fundamental level. Let me here just breezily offer the suggestion that the laws of causation assert the identity of one state of the world with a later state of the world: so for example to say that the world featured me striking a match in certain appropriate conditions at one moment is equivalent (under these laws) to saying there was fire at a slightly later moment. Now if that is true, the only possible causes of the existence of the cosmos at its first moment are either its non-existence at any earlier moment or its own existence at that first moment (if you allow simultaneous causation). I think this says the universe is either necessarily gratuitous or gratuitously necessary, but I'll leave it with you there for now.

Why do the contents of the world seem so arbitrary and random? I suggest there are two reasons. First, the ongoing transcendence which drives the universe is nomic as well as ontic. It's not just that there's more stuff, there are more, and more complex, underlying laws. Our view of the long-term past and future is therefore obscured: the ancient universe was not just physically smaller but metaphysically impoverished or cramped, too, and long-term extrapolations are systematically thrown off by this. If we could understand the process properly, it may be that things would look less random – though I grant that for the moment this must be an optimistic article of faith rather than a rigorously deduced conclusion.

Let me just pour myself a bit of a digression here on the nature of the laws of nature. It is common to speak of the laws of nature or the laws of physics although this is clearly a metaphor, and a very old one. Few people, I would guess, suppose that the laws of nature are literally written down in some cosmic text and enforced by angelic police officers – although it is not uncommon to suppose that the mathematics we use to describe the world is actually what controls it, which I think may be a similar error. So what are these laws? One problem for us is that not only are they not written down in any cosmic text, they're not written down on paper in earthly texts either: so far as I know, no-one has ever set down a comprehensive list of the Laws of Nature. Physics textbooks set out a number of laws, indeed, but these tend to be the non-obvious ones, rather in the way that early dictionaries included only 'hard' words. The nearest thing to a full statement might be in those efforts to produce a Naïve Physics that ended up (in my opinion) producing something that actually struck the normal mind as far weirder than mere Newtonian physics; but they were explicitly setting out a misconceived version of the laws.

It's consequently hard to feel assured that all relevant examples have been covered: but again I will cut boldly to the chase and suggest that all laws of nature are in fact laws of conservation. They can all be reconstituted as assertions of the continued existence of an underlying entity in different cases, usually at different times. Certainly it seems that any law which can be stated in

the form of an equation must be of this kind, because the equation of x with y essentially tells us that the quantity of underlying z of which they are both expressions remains the same whichever form we take it in. Laws which assert, or contain, a constant, clearly state the existence of a continuing fundamental entity in even clearer form.

If that's true, then perhaps the laws of nature could be restated as a list of existential assertions (though some of the entities whose existence is asserted would be a little unfamiliar), which with a list of values would give us a comprehensive anomic account of the world.

Be that as it may, and getting back to the main point, it is at any rate clear that besides any confusion arising from any nomic evolution which may be going on, certain features of the world arise out of the operation of causality over time. Now it could be that time itself is actually constituted by the ontic growth of the universe (the steady drip of extra stuff providing the steady tick of the moments); but in any case that growth clearly requires time. It may be that some of the deep constant features of the universe are sustained in their existence directly by the same inscrutable principle which caused the universe to come into existence in the first place; but others definitely depend on a long stream of complex causality, and this is surely where the haecceity primarily comes in. We could say that everything is necessary, but that while some things are necessary in the light of metaphysics, others are necessary in the light of history.

To restate that: it may well be that the universe in which we find ourselves is actually the only possible one, and the product of a necessary ontological (and nomological) evolution; but the necessity of the details is both obscured from us by the nature of the evolution itself and also genuinely different from the direct necessity of the underlying features in that it derives its necessity through causation over time.

That's why actual experience and actual qualia seem so strange and so difficult to capture theoretically. I hope that makes sense of some kind and perhaps appeals to some degree: I'm away now to sleep it off.

Moral Machines – Thinking Otherwise

2 September 2012

David Gunkel of NIU has produced a formidable new book (via) on the question of whether machines should now be admitted to the community of moral beings.

He lets us know what his underlying attitudes are when he mentions by way of introduction that he thought of calling the book A Vindication of the Rights of Machines, in imitation of Mary Wollstonecraft. Historically Gunkel sees the context as one in which a prolonged struggle has gradually extended the recognised moral domain from being the exclusive territory of rich white men to the poor, people of colour, women and now tentatively perhaps even certain charismatic animals (I think it overstates the case a bit to claim that ancient societies excluded women, for example, from the moral realm altogether: weren't women like Eve and Pandora blamed for moral failings, while Lucretia and Griselda were held up as fine examples of moral behaviour - admittedly of a rather grimly self-subordinating kind? But perhaps I quibble.) Given this background the eventual admission of machines to the moral community seems all but inevitable; but in fact Gunkel steps back and lowers his trumpet. His more modest aim, he says, is like Socrates simply to help people ask better questions. No-one who has read any Plato believes that Socrates didn't have his answers ready before the inquiry started, so perhaps this is a sly acknowledgement that Gunkel too, thinks he really knows where this is all going.

For once we're not dealing here with the Anglo-Saxon tradition of philosophy: Gunkel may be physically in Illinois, but intellectually he is in the European Continental tradition, and what he proposes is a Derrida-influenced deconstruction. Deconstruction, as he concisely explains, is not destruction or analysis or debunking, but the removal from an issue of the construction applied heretofore. We can start by inverting the normal understanding, but then we look for the emergence of a new way of 'thinking otherwise' on the topic which escapes the traditional framing. Even the crustiest of Anglo-Saxons ought to be able to live with that as a modus operandi for an enquiry.

The book falls into three sections: in the first two Gunkel addresses moral agency, questions about the morality of what we do, and then moral patiency, about the morality of what is done to us. This is a sensible and useful division. Each section proceeds largely by reportage rather than argument, with Gunkel mainly telling us what others have said, followed by summaries which are not really summaries (it would actually be very difficult to summarise the multiple, complex points of view explored) but short discussions moving the argument on. The third and final section discusses a number of proposals for 'thinking otherwise'.

On agency, Gunkel sets out a more or less traditional view as a starting point and notes that identifying agents is tricky because of the problem of 'other minds': we can never be sure whether some entity is acting with deliberate intention because we can never know the contents of another mind. He seems to me to miss a point here; the advent of the machines has actually transformed the position. It used to be possible to take it for granted that the problem of

other minds was outside the scope of science, but the insights generated by AI research and our ever-increasing ability to look inside working brains with scanners mean that this is no longer the case. Science has not yet solved the problem, but the idea that we might soon be able to identify agents by objective empirical measurement no longer requires reckless optimism.

Gunkel also quotes various sources to show that actual and foreseen developments in AI are blurring the division between machine and human cognition (although we could argue that that seems to be happening more because the machines are getting better and moving into marginal territory than because of any fundamental flaws in the basic concepts). Instrumentalism would transfer all responsibility to human beings, reducing machines to the status of tools, but against that Gunkel quotes a Heideggerian concept of machine autonomy and criticisms of inherent anthropocentrism. He gently rejects the argument of Joanna Bryson that robots should be slaves (in normal usage I think slaves have to be people, which in this discussion begs the question). He rightly describes the concept of personhood as central and points out its links with consciousness, which, as we can readily agree, throws a whole new can of worms into the mix. All in all, Gunkel reasonably concludes that we're in some difficulty and that the discussion appears to 'lead into that kind of intellectual cul-de-sac or stalemate that Hegel called a "bad infinite."'

He doesn't give up without considering some escape routes. Deborah Johnson interestingly proposes that although computers lack the metal states required for proper agency, they have intentionality and are therefore moral players at some level. Various others offer proposals which lower the bar for moral agency in one way or another; but all of these are in some respects unsatisfactory. In the end Gunkel thinks we might do well to drop the whole mess and try an approach founded on moral patiency instead.

The question now becomes, not should we hold machines responsible for their actions, but do we have a moral duty to worry about what we do to them? Gunkel feels that this side of the question has often been overshadowed by the debate about agency, but in some ways it ought to be easier to deal with. An interesting proposition here is that of a Turing Triage Test: if a machine can talk to us for a suitable time about moral matters without our being able to distinguish it from a human being, we ought to give it moral status and presumably, not turn it off. Gunkel notes reasonable objections that such a test requires all the linguistic and general cognitive capacity of the original Turing test simply in order to converse plausibly, which is surely asking too much. Although I don't like the idea of the test very much, I think there might be ways round these objections if we could reduce the interaction to multiple-choice button-pushing, for example.

It can be argued that animals, while lacking moral agency, have a good claim to be moral patients. They have no duties, but may have rights, to put it another way. Gunkel rightly points here to the Benthamite formulation that what matters is not whether they can think, but whether they can suffer; but he notes considerable epistemological problems (we're up against Other Minds again). With machines the argument from suffering is harder to make because hardly anyone believes they do suffer: although they may simulate emotions and pain, it is most often agreed that in this area at least Searle was right that simulations and reality are poles apart. Moreover it's debatable whether bringing animals into a human moral framework is an unalloyed benefit or to some extent simply the reassertion of human dominance. Nevertheless some would go still further and Gunkel considers proposals to extend ethical status to plants, land, and artefacts. Information Ethics essentially completes the extension of the ethical realm by excluding nothing at all.

This, then is one of the ways of thinking otherwise - extending the current framework to include non-human individuals of all kinds. But there are other ways: one is to extend the individual: a number of influential voices have made the case in recent years for an extended conception of consciousness, and that might be the most likely way for machines to gravitate within the moral realm - as adjuncts of a more broadly conceived humanity.

More radically, Gunkel suggests we might adopt proposals to decentre the system; instead of working with fixed Cartesian individuals we might try to grant each element in the moral system the rights and responsibilities appropriate to it at the time (I'm not sure exactly how that plays out in real situations); or we could modify and distribute our conception of agency. There is an even more radical possibility which Gunkel clearly finds attractive in the ethics of Emmanuel Levinas, which makes both agency and patiency secondary and derived while ethical interactions become primary, or to put it more accurately:¶

> *The self or the ego, as Levinas describes it… becomes what it is as a by-product of an uncontrolled and incomprehensible exposure to the face of the Other that takes place prior to any formulation of the self in terms of agency.*

I warned you this was going to get a bit Continental - but actually making acts define the agent rather than the other way about may not be as unpalatably radical as all that. He clearly likes Levinas' drift, anyway, and perhaps even better Silvia Benso's proposed 'mash-up' which combines Levinasian non-ethics with Heideggerian non-things (tempting but a little unfair to ask what kind of sense that presumably makes).

Actually the least appealing proposal reported by Gunkel, to me at least, is that of Anne Foerst, who would reduce personhood to a social construct which we assign or withhold: this seems dangerously close to suggesting that say, concentration camp guards can actually withdraw real moral patienthood from their victims and hence escape blame (I'm sure that's not what she means).

However, on the brink of all this heady radicalism Gunkel retreats to common sense. At the beginning of the book he suggested that Descartes could be seen as 'the bad guy' in his reduction of animals to the status of machines and exclusion of both from the moral realm; but perhaps after all, he concludes, we are in the end obliged to imitate Descartes' provisional approach to life, living according to the norms of our society while the philosophical issues resist final resolution. This gives the book a bit of a dying fall and cannot but seem a bit of a cop-out.

Overall, though, the book provides a galaxy of challenging thought to which I haven't done anything like justice and Gunkel does a fine job of lucid and concise exposition. That said, I don't find myself in sympathy with his outlook. For Gunkel and others in his tradition the ethical question is essentially political and under our control: membership of the moral sphere is something that can be given or not given rather like the franchise. It's not really a matter of empirical or scientific fact, which helps explain Gunkel's willingness to use fictional examples and his relative lack of interest in what digital computers actually can and cannot do. While politics and social convention are certainly important aspects of the matter, I believe we are also talking about real, objective capacities which cannot be granted or held back by the fiat of society any more than the ability to see or speak. To put it in a way Gunkel might find more congenial: when ethnic minorities and women are granted equal moral status, it isn't simply an

arbitrary concession of power, the result of a social tug-of-war but the recognition of hard facts about the equality of human moral capacity.

Myself I should say that moral agency is very largely a matter of understanding what you are doing; an ability to allow foreseen contingencies to influence current action. This is something machines might well achieve: arguably the only reason they haven't got it already is an understandable human reluctance to trust free-ranging computational decision-making given an observable tendency for programs to fail more disastrously and less gracefully than human minds.

Moral patienthood on the other hand is indeed partly a matter of matter of the ability to experience pain and other forms of suffering, and that is problematic for machines; but it's also a matter of projects and wishes, and machines fall outside consideration here because they simply don't have any. They literally don't care, and hence they simply aren't in the game.

That seems to leave me with machines that I should praise and blame but need not worry about harming: should be good for a robot butler anyway.

Libet Plays Ball

16 Septemberr 2012

Recent research at UCL provides corroboration for the claim that elite sports players often feel they have plenty of time to think about how they address a ball which is actually coming at them so fast that normal human beings probably shouldn't even see it.

This effect, a perceived slowing down of time, is the kind of thing which I think would once have been shrugged off as unresearchable, outside the scope of proper science: how can we tell how things seem to people? These days we're bolder, and the team in this case came up with an experiment in which some subjects were asked to tap a screen while others were merely asked to observe. Those who were asked to prepare action reported feeling the time they had available seemed longer.

The finding makes a certain obvious sense in that being granted extra mental time when a particularly crucial task is coming up is bound to be helpful. But it's also a little puzzling: if the brain is capable of giving us more time, why doesn't it do so routinely? It would never be a bad thing to have more time for reflection; if the facility only gets turned on in these special circumstances there must either be some downside or cost which makes the brain ration its use, or something else is wrong.

It's not altogether implausible that revving up our mental processes might have an energy cost, or that our neurons might be able to speed up a bit only temporarily; but as always with consciousness there's a deeper issue. Did mental processing actually speed up, or did it simply feel as if it did; or indeed, was the time merely remembered as passing more slowly? There's an assumption here that when we are playing baseball or tennis control is fully conscious, but actually it's far from clear that the deliberative level of thought has much to do with it - it generally seems more a matter of gut reaction than strategic debate.

Another complicating factor is that we don't actually perceive the passage of time directly, in the way we can observe spatial sizes. There is no special organ devoted to measuring time, and it seems likely that if we do have to assess time (we're pretty bad at it- try telling when two minutes have passed without looking at a clock) we pick up on several different indirect clues.

It may well be that one way the brain works out how much time has passed is by counting the number of 'events' it remembers and assuming that the gap between each of them is about the same. 'Events' would be mental ones, which explains why outwardly uneventful time may seem to pass very slowly - because we keep checking our watches or asking ourselves how much longer, so that there is a steady stream of mental acts, whereas when we're caught up in something interesting there are none of these mental markers put down and time seems to fly.

It seems highly plausible to me that when we're asked to prepare mentally for a given act (whether tapping the screen or hitting a perfect drive to the boundary) a sequence of this kind is set up, in which we keep asking ourselves 'am I going to do it now?' or 'is it time yet?'. If that's

true the chances are the impression of having more time to react is actually an illusion, though it may reflect a genuinely improved state of readiness.

I couldn't help reflecting that the conditions of these experiments rather resembled those in the famous series of experiment by Benjamin Libet which seemed to show that a decision to move one's hand at a given moment had actually been taken significantly before that same decision entered consciousness. Libet's subjects were asked to get ready to act in just the sort of way which might have led to the kind of time-dilation effect considered in the UCL research. Libet's ingenious system for measuring the time of decision, with subjects reporting the position of a clock at the vital moment, should be fairly well proofed against subjective errors: but could it be that a sense of time slowing down caused Libet's subjects to delay slightly in reporting their decision?

Just a thought.

Oh Phi!

1 October 2012

Giulio Tononi's Phi is an extraordinary book. It's heavy, and I mean that literally: presumably because of the high quality glossy paper, it is noticeably weighty in the hand; not one I'd want to hold up before my eyes for long without support, though perhaps my wrists have been weakened by habitual Kindle use.

That glossy paper is there for the vast number of sumptuous pictures with which the book is crammed; mainly great works of art, but also scientific scans and diagrams (towards the end a Pollock-style painting and a Golgi-Cox image of real neurons are amusingly juxtaposed: you really can barely tell which is which). What is going on with all this stuff?

My theory is that the book reflects a taste conditioned by internet use. The culture of the World Wide Web is quotational and referential: it favours links to good stuff and instant impact. In putting together a blog authors tend to gather striking bits and pieces rather the way a bower bird picks up brightly coloured objects to enhance its display ground, without worrying too much about coherence or context. (If we were pessimistic, we might take this as a sign that our culture, like classical Greek culture before it, is moving away from the era of original thought into an age of encyclopedists; myself I'm not that gloomy - I think that however frothy the Internet may get in places it's all extra and mostly beneficial.) Anyway, that's a bit what this book is like; a site decorated with tons of 'good stuff' trawled up from all over, and in that slightly uncomplimentary sense it's very modern.

You may have guessed that I'm not sure I like this magpie approach. The pictures are forced into a new context unrelated to what the original artist had in mind, one in which they jostle for space, and many are edited or changed, sometimes rather crudely (I know: I should talk, when it comes to crude editing of borrowed images - but there we are). The choice of image skews towards serious art (no cartoons here) and towards the erotic, scary, or grotesque. Poor Maddalena Svenuta gets tipped on her back, perhaps to emphasise the sexual overtones of the painting - although they are unignorable enough in the original orientation. This may seem to suggest a certain lack of respect for sources and certainly produces a rather indigestible result; but perhaps we ought to cut Tononi a bit of slack. The overflowing cornucopia of images seems to reflect his honest excitement and enthusiasm: he may, who knows, be pioneering a new form which we need time to get used to; and like an over-stuffed blog, the overwhelming gallimaufry is likely here and there to introduce any reader to new things worth knowing about. Besides the images the text itself is crammed with disguised quotes and allusions. Prepare to be shocked: there is no index.

I'm late to the party here. Gary Williams has made some sharp-eyed observations on the implicit panpsychism of Tononi's views; Scott Bakker rather liked the book and the way some parts of Tononi's theory chimed with his own Blind Brain theory (more on that another time, btw). Scott, however, raised a 'quibble' about sexism: I think he must have in mind this hair-raising sentence in the notes to Chapter 29:¶

> *At the end, Emily Dickinson saves the day with one of those pronouncements that show how poets (or women) have deeper intuition of what is real than scientists (or men) ever might: internal difference, where all the meanings are.*

Ouch, indeed: but I don't think this is meant to be Tononi speaking.

The book is arranged to resemble Dante's Divine Comedy in a loose way: Galileo is presented as the main character, being led through dreamlike but enlightening encounters in three main Parts, which in this case present in turn, more or less, the facts about brain and mind - the evidence, the theory of Phi, and the implications. Galileo has a different guide in each Part: first someone who is more or less Francis Crick, then someone who is more or less Alan Turing, and finally for reasons I couldn't really fathom, someone who is more or less Charles Darwin (a bit of an English selection, as the notes point out); typically each chapter involves an encounter with some notable personality in the midst of an illuminating experience or experiment; quite often, as Tononi frankly explains, one that probably did not feature in their real lives. Each chapter ends with notes that set out the source of images and quotations and give warnings about any alterations: the notes also criticise the chapter, its presentation, and the attitudes of the personalities involved, often accusing them of arrogance and taking a very negative view of the presumed author's choices. I presume the note writer is, as it were, a sock puppet, and I suppose this provides an entertaining way for Tononi to voice the reservations he feels about the main text, backing up the dialogues within that text with a kind of one-sided meta-textual critique.

Dialogue is a long-established format for philosophy and has certain definite advantages: in particular it allows an author to set out different cases with full vigour without a lot of circumlocution and potential confusion. I think on the whole it works here, though I must admit some reservation about having Galileo put into the role of the naive explorer. I sort of revere Galileo as a man whose powers of observation and analysis were truly extraordinary, and personally I wouldn't dare put words into his mouth, let alone thoughts into his head: I'd have preferred someone else: perhaps a fictional Lemuel Gulliver figure. It makes it worse that while other characters have their names lightly disguised (which I take to be in part a graceful acknowledgement that they are really figments of Tononi) Galileo is plain Galileo.

Why has Tononi produced such an unusual book? Isn't there a danger that this will actually cause his Integrated Information Theory to be taken less seriously in some quarters? I think to Tononi the theory is both exciting and illuminating, with the widest of implications, and that's what he wants to share with us. At times I'm afraid that enthusiasm and the torrent of one damn thing after another became wearing for me and made the book harder to read: but in general it cannot help but be engaging.

The theory, moreover, has a lot of good qualities. We've discussed it before: in essence Tononi suggests that consciousness arises where sufficient information is integrated. Even a small amount may yield a spark of awareness, but the more we integrate, the greater the level of consciousness. Integrated potential is as important as integrated activity: the fact that darkness is not blue and not loud and not sweet tasting makes it, for us, a far more complex thing that it could ever be to an entity that lacked the capacity for those perceptions. It's this role of absent or subliminal qualities that make qualia seem so ineffable.

This makes more sense than some theories I've read but for me it's still somewhat problematic. I'm unclear about the status of the 'information' we're integrating and I don't really understand

what the integration amounts to, either. Tononi starts out with information in the unobjectionable sense defined by Shannon, but he seems to want it to do things that Shannon was clear it couldn't. He talks about information having meaning when seen from the inside, but what's this inside and how did it get there? He says that when a lot of information is aggregated it generates new information - hard to resist the idea that in the guise of 'new information' he is smuggling in a meaningfulness that Shannonian information simply doesn't have. The suggestion that inactive bits of the system may be making important contributions just seems to make it worse. It's one thing for some neural activity to be subliminal or outside the zone of consciousness: it's quite different for neurons that don't fire to be contributing to experience. What's the functional difference between neurons that don't fire and those that don't exist? Is it about the possibility that the existing ones could have fired? I don't even want to think about the problems that raises.

I don't like the idea of qualia space, another of Tononi's concepts, either. As Dennett nearly said, what qualia space? To have an orderly space of this kind you must be able to reduce the phenomenon in question to a set of numerical variables which can be plotted along axes. Nobody can do this with qualia; nobody knows if it is even possible in principle. When Wundt and his successors set out to map the basic units of subjective experience, they failed to reach agreement, as Tononi mentions. As an aspiration qualia space might be reasonable, but you cannot just assume it's OK, and doing so raises a fear that Tononi has unconsciously slipped from thinking about real qualia to thinking about sense-data or some other tractable proxy. People do that a lot, I'm afraid.

One implication of the theory which I don't much like is the sliding scale of consciousness it provides. If the level of consciousness relates to the quantity of information integrated, then it is infinitely variable, from the extremely dim awareness of a photodiode up through flatworms to birds, humans and - why not - to hypothetical beings whose consciousness far exceeds our own. Without denying that consciousness can be clear or dim, I prefer to think that in certain important respects there are plateaux: that for moral purposes, in particular, enough is enough. A certain level of consciousness is necessary for the awareness of pain, but being twice as bright doesn't make my feelings twice as great. I need a certain level of consciousness to be responsible for my own actions, but having a more massive brain doesn't thereafter make me more responsible. Not, of course, that Tononi is saying that, exactly: but if super-brained aliens land one day and tell us that their massive information capacity means their interests take precedence over ours, I hope Tononi isn't World President.

All that said, I ought to concede that in broad terms I think it's quite likely Tononi is right: it probably is the integration of information that gives rise to consciousness. We just need more clarity about how - and about what that actually means.

Glockner's Isthmus

7 October 2012

Consciousness, as a subject, is nothing if not multidisciplinary. We have noted once or twice before that besides all the different groups of scientists and philosophers, novelists are strongly interested; in fact it's not much of a stretch to claim that novel-writing is itself an exploration of human consciousness. If so, it's an exploration that is generally pursued separately from the academic investigation, but the occasional linkages and crossovers are always interesting. Did Henry James's works influence his brother William's psychology? Alas, we'll never know exactly, but William's concept of the stream of consciousness had a massive impact on novelists.

Novels are certainly extended, virtuoso exercises in theory of mind. Novelists use their own understanding of other minds to evoke in the consciousness of readers an intense engagement with the purely imagined mental lives of their characters, and it's all achieved purely through print on a page; a remarkably complex feat to pull off. As a result novelists arguably have a better appreciation than most cognitive scientists of just how complicated the actual experience of consciousness really is.

When you step back and look at it, the everyday experience of consciousness really is complicated, isn't it? From moment to moment a gamut of miscellaneous sensory and proprioceptive impressions, memories, fragments of explicit speech, emotions and desires all float in and out of the several levels of conscious, unconscious, and half-conscious thought that run in parallel, with some in the centre or the penumbra of a focus of attention that may at times not be present at all, and others in one or more levels of background; all conditioned by prevailing but variable states of being calm, exalted, curious, depressed, drunk; the whole thing pushed along at times by perceptions, impulses and predispositions emerging unobtrusively from silent faculties outside the pale of true mentality. All of that, or the gist of it, has to represented in a linear text in such a way as to bring about the required 'willing suspension of disbelief' in the reader, or indeed more than that, a positive, empathetic identification.

The basic technique, I suppose, is to present a text which seems to record someone recounting his own experiences. One option for complete realism is to present the text as letters or diary entries. Or like Plato, we could present two friends meeting by chance, one of whom then relates a memorable conversation Socrates once had. Already there another narrator has slipped unobtrusively into the background - the unmentioned person - Plato? - who must be telling us what the two friends said. This shadowy personage can, if we wish, do the whole job of narration for us, becoming the mysteriously omniscient narrator we have become used to in standard storytelling. The range of options that open up thereafter is wide: first person, third person, past tense, present historic, dialogue, unreliable narrator; the sophisticated free indirect style which allows us to slip smoothly between the description of physical events and mental ones, and more radical techniques like the aforementioned stream of consciousness;

lately novelists have been able to use, or had to contend with, the narrative conventions and techniques drilled into readers by films and graphics.

It has been suggested more than once that the self and our conscious experience come from the brain spinning a narrative. There may well be some truth in that, but in the light of all the foregoing I think we can see that it raises as many questions as it answers. What kind of narrative? Visual? Textual? First person? Are we allowed flashbacks and jump cuts? Far from narrative being a simple primitive, something well understood which we can use to explain consciousness and the self, it seems to be one of consciousness's most complicated and surprising final products.

If we under-rate the complexity of the task facing novelists, they may be partly to blame; it seems evident that to a great extent our view of our own conscious lives has been shaped by reading all those novels. We sort of expect our inner lives to resemble the neatened-up depictions assigned to characters in books, and that's how we tend to think of them. More than that, and perhaps rather scarily, we may suspect that our mental lives have actually been changed and in some degree shaped by those expectations.

So with all that by way of a build-up - what do we think of A Possible Life, Sebastian Faulks' latest?

Most reviews have suggested the book isn't really a novel at all, but a collection of short stories. Certainly it looks that way, although the separate stories have linkages: shared or duplicated experiences, common characters or objects. It seems natural to suppose that Faulks owes something to David Mitchell in this respect; but whereas Mitchell's Cloud Atlas hinted at reincarnation Faulks seems, if we can trust the thoughts of one character which close the book, to be offering a more radical thesis: let's come back to that, though.

The individual stories are pretty good and sketched out with a light but authoritative touch: I would single out a nineteenth-century one which I thought was remarkable. It's not easy to depict Victorians from the poorer classes for a number of reasons (not least the looming presence of Dickens) but I thought this story rang true on every level. However, for Conscious Entities the most interesting one concerns two scientists who, in the near future, solve the mystery of consciousness. It's clear that Faulks has done his research into the current state of the field, and we can spot the influence of Damasio and others in the sketch he offers us (again pretty convincing) of how the breakthrough arrives. The critical enabler proves to be, guess what, an advance in scanning technology, and the breakthrough is preceded by the discovery of Glockner's Isthmus: a linkage which momentarily binds sensory data with input from the internal organs, providing a sense of self which is apparently common to humans and some other animals. This, however, provides only base-level consciousness: human-style self-awareness relies on the Rossi-Duranti Loop which connects the Isthmus with episodic memory.

This is a novel, not an academic paper: otherwise, we might ask for a bit more detail on how a link and a loop are sufficient to do the ontological and cognitive work being asked of them here; but it's not a bad thesis and the gist is clear: a sense of self from the guts and a context from memory. We are in fact dealing with a system designed to insert a self into a narrative. Proust, a literary critic suggests in the story, has been vindicated.

One incidentally interesting feature of Faulks' system is that the Rossi-Duranti Loop only works now and then, with the implication that full consciousness is only switched on at particularly reflective moments. I think this mainly means that we operate largely on autopilot, with

conscious decision-making and creativity reserved for when they're needed or invoked, but it could suggest that most of the time we are philosophical zombies, without qualia; living examples of a thought-experiment we've mentioned before. Faulks makes the Loop a recent development, so that, in a sort of echo of Julian Jaynes, those old cave men and people from early times were not really conscious, in spite of their splendid paintings.

Our scientists are helped by the chance arrival of a subject who, rather like Phineas Gage, has had a rod driven through his head: in this case a kebab skewer. On Kebab Man the effect is to turn his Loop permanently on, but Faulks, alas, resists the urge to show or tell us what this is like (Marvellous? Exhausting? If this were SF, Kebab Man might become a super hero or a genius like those in Ted Chiang's Understand).

Has the arrival of a naturalistic explanation for consciousness destroyed something important? Some, in the novel, think it has. Elena Duranti, one of the scientists responsible for discovering the Loop, has a childhood friend (or rather more than a friend) called Bruno who is a gifted story-teller and later an author(who talks about the narrative of their lives and the question, Copperfield-like, of being the hero of one's own story). Angrily, towards the end, he tells her she has proved we don't really exist and thereby brought despair: "I don't think that's what Beatrice and I proved." Elena says.

What did they prove? Well, I think the radical thesis I mentioned above, which I take to be what Faulks is putting forward, is that the separate experiences of our life are not inherently unique to us, but shared or potentially shared: that our personal identity is only a loose linkage from a wider pool of shared experience, of which we are all part. No real need to fear death, then, because it follows that ultimately "we're all in this thing, like it or not, for ever." It is sometimes suggested that individual human consciousness emerges out of a general cosmic mind: this is the cosmic mind, nor am I out of it, Faulks might say.

If that is indeed the thesis of the book, then we can see that it could only be illustrated by apparently separate stories which at some level and in places are united, so the contention that it's less a novel than a collection sort of misses the point. At any rate, I recommend it to anyone who is interested in this stuff.

The Cambridge Declaration

14 October 2012

At the Francis Crick Memorial Conference back in July the participants signed a Declaration affirming that animals are conscious. The key passage reads: ¶

"The absence of a neocortex does not appear to preclude an organism from experiencing affective states. Convergent evidence indicates that non-human animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviors. Consequently, the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Nonhuman animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates."

Actually, you'll notice that that paragraph crams on the brakes an inch short of saying that animals are conscious. It says they have various substrates and that it doesn't seem to be impossible that they should have 'affective states'. More boldly it says they have the capacity to exhibit intentional behaviours.

We could certainly do with some clarity on this issue - but why are scientists issuing a Declaration? Don't they just publish papers and leave the results to speak for themselves? There's a clear intention that the Declaration should have political impact, at least encouraging protection of animals, perhaps lending support to the idea of animal rights, possibly even giving a boost to veganism if we assume Philip Low was serious. (In fact I suppose we might have to go further and intervene to prevent carnivores of other species pursuing their traditional diet.)

Whatever we feel about that, it takes us into contentious territory, and I think there are in any case two other factors which make discussion in this area especially complicated. One is that there is a hidden dispute here about standards of proof. At least three standards are, as it were, in play. There's the ordinary/political standard of proof, which requires something to be shown to be true clearly enough that we believe it and generally expect others to believe it. There's the scientific standard of proof, which requires more rigorous evidence, reproducibility of results, and endorsement by peer review. Then there's the philosophical standard of proof, which requires that no rational being who understands the case could possibly think otherwise, which actually makes it tough to prove that animals even exist, never mind what their mental states are. All of these different standards have their place, but discussion of animal consciousness is handicapped by differences and misunderstandings about which to invoke in which contexts.

Animal consciousness is also a peculiarly emotive subject. Our conclusions may be very strongly affected by feelings of empathy, and indeed there is an argument that empathy is an

appropriate response. Although I've never heard this particular argument made explicitly, it can be argued legitimately that even if animals are not conscious, we are better people if we let our empathy cause us to believe they are, because empathy is such a vital human quality it even trumps strict truth in this case. It's better to waste some sympathy on creatures that don't really deserve it than risk remaining cold to those that do, in other words. At any rate some may feel that those who deny that animals have feelings at all (step forward, Descartes) are deficient in perception or even perhaps in moral sense, a feeling that sceptics on their side are not apt to welcome.

So discussion is potentially difficult, and so it is with all due caution I say I think the Declaration is probably a bad idea overall; and that there are three particular problems with it.

The first problem is that it doesn't seem well-motivated scientifically, by which I mean it doesn't seem to be inspired by any definite scientific result. The preliminary passages of the Declaration do allude to various pieces of research, but the basic case is that animals have much of the same brain hardware as human beings, and also exhibit the right sorts of behaviour. We've known that for a long time. If some scientific Rubicon has been crossed recently, I missed it and I can't see it set out here. To make matters worse the Declaration seems to come close to a clunking logical error, along the lines of: other areas than the neocortex are involved in having feelings; animals have those other areas, therefore animals have feelings. That wouldn't work: you could as well argue that: other organs than the eye are involved in seeing; people whose eyes have been gouged out have those other organs; therefore people whose eyes have been gouged out can see. You can't really dismiss the neocortex that easily.

The second problem is that we're not told clearly which animals are conscious, and no allowance is made for different levels of consciousness in different species. On the face of it it seems the claim is that insects are just as conscious as anyone else (so those classic observations of stereotyped behaviour in sphex wasps must somehow have been wrong). In fact Christof Koch seems to be saying everything is conscious. That's too extreme to be plausible and too vague to be helpful, especially in a context where we're implicitly thinking about political rights and moral status. This is by no means an academic problem in a world where, for one thing, experimental animals may be sacrificed to help save human lives; if we think all beings have an equal share of consciousness there will certainly be no more animal experiments, and if we get it wrong either way that has serious consequences.

The third problem is that we're not told what kind of consciousness. The Declaration seems mainly to be about whether animals feel pain, but it makes no distinction between a capacity for suffering and a capacity for rational planning (or indeed self-awareness). These are radically different things, and we might well find it much more likely that animals have the former than the latter.

This is important in two respects. First, if animals have feelings, then we might want to enact laws protecting them from pain; but we won't feel inclined to give them rights unless we think they have the kind of intentional thought that would allow them to exercise those rights. To have legal rights you need to be the kind of entity that might go to court to enforce them, and presumably the kind of entity which also has duties and gets punished for infractions. Actually, that's not quite true; we might still want to give animals the special kind of rights that children or comatose patients have, which need to be exercised on their behalf by others and entail no obligations; but even if we do that it's important to keep track of what legal/moral status we're actually awarding.

Second, the presence or absence of rational conscious thought might affect our view of how important feelings are. It might be that some animals can indeed feel pain in a basic way, but that without human-style neocortex-driven awareness their pains are only like dream ones, or like pains we instantly forget (some people take the view that a forgotten agony doesn't matter; and indeed surgeons use mnestics, which make you forget the pain, almost interchangeably with true anaesthetics, which actually stop you feeling it in the first place). Even if animals feel pain it might still be that their pains are of lesser value or even of no account at all; and the presence or absence of explicitly entertained thoughts, projects and desires might be relevant.

There's an awful lot more to be said about these issues, quite a lot of it on the Declaration's side, but that brings me to the overall problem. What is the Declaration for, and what will it be used for? I presume the aspiration is that it will be cited wherever the treatment of animals is an issue. Its function is surely meant to be to give clarity, but to do it by curtailing argument; to end discussion. That's certainly liable to be the way it is used, at any rate. I don't like that at all; there are cases where it's legitimate to try to close down an area of discussion, but I don't think this is one; the attempt is at best radically premature, at worst profoundly unhelpful. If anything we should be promoting discussion. Wouldn't it have been better, and more appropriate to Crick's memory, if the participants in the Conference had, in that spirit, published a list of good questions?

Near Death: A View From Hell

22 Octobee 2012

My dear Wormwood,

How delightful that after so long you should again be seeking my counsel! I remain, of course, your ever-affectionate Uncle; but what a tizzy you seem to be in! I suppose it is understandable. Here you are, a junior devil - still only a junior devil, alas - and you find that the soul you have been set to ensnare has suddenly become a leading champion of the Enemy! Eben Alexander, a neurologist of good standing and therefore surely our rightful prey, has had a near-death experience, and on his recovery has published a book sensationally claiming that he now knows the truth of the afterlife at first hand. Why me, you complain in whining tones, why does it have to be me that gets the Celebrity Christian?

Please compose yourself. As a matter of fact the situation is nothing like as bad as you suppose; in fact I will venture to say that it is excellent. In the first place you talk as though Dr Alexander had undergone a miraculous conversion, but in fact is it not the case that he was always a declared Christian? Of course, we should prefer him not to be a Christian at all, but if it must be so then a Celebrity Christian, I assure you, is the very best kind he could be. Consider, to begin with, the opportunities for that most useful sin, spiritual vanity. I say opportunities, but are we not in fact dealing with a sin which is already fully realised? The symptoms of vanity are bad enough in those people who merely insist on telling one about their dreams; this fellow believes his are a Divine revelation which we must all read about! I hardly think Dr Alexander, in the quiet of his own mind, can suppose himself to be much less than a saint - and then which of the saints could claim to have the further distinction of being a proper scientist? A neurosurgeon, to boot! He surely feels that he has been called to a high and lonely eminence. Apart from its inherently damnable qualities, this vanity will encourage a misplaced feeling of certainty and divert his attention away from the very area - his own failings and imperfections - which most need his attention. And of course, it is always especially delightful when a man's religion is the very thing that helps drag him Hellwards.

In passing we may note that while firm faith is highly adverse to our cause, a feeling of scientific certainty about God and Heaven is very helpful to us. The humans forget that faith can only exist where there is doubt (otherwise why would the Enemy leave them to work things out for themselves in that perverse manner of his?) and forget that if they behave well merely in order to get into a scientifically established Heaven, their acts are self-interested and lack true virtue. Our Lord Below once pointed out to the Enemy how well this all works for us: look, he said; to be virtuous the humans must not be acting for reward, so only someone who doesn't believe in Heaven can really be good. Only the atheists have an opportunity of disinterested virtue - and obviously only the believers have faith; so between the two tests virtually no-one can qualify for Paradise. No wonder, he said, that the Pearly Gates stand almost unused, disconsolate angels telling each other hopefully that surely this year a couple of people will get in, while all the time the mouth of Hell gapes and swallows, gapes and swallows, a thousand miles wide every time. I'm not asking you to concede defeat in any humiliating way, Our Father Below explained to the

Enemy, all I'm saying is, be realistic - let's get round a table on this one and see what we can work out? I regret to say, dear nephew, that these very rational and reasonable overtures were summarily rejected.

The second splendid thing about a Celebrity Christian is the scope for humbug. It's virtually impossible for these people to be completely honest. One of Dr Alexander's claims to a special status for his revelation is that, uniquely, it was scientifically established that his neocortex was entirely inactive throughout the period of his absence; established by neurological investigations and CT scans. Now as a neurosurgeon Dr Alexander must be well aware that even fMRI scans only measure neuronal activity via the proxy of blood flow (albeit it's a fairly good proxy), and that CT scans don't measure it at all, merely looking for damage and possible bleeding. The 'neurological tests' we may confidently take to be little more than the observation that he was unresponsive while unconscious. None of this would actually prove that his neocortex was completely inactive. Even if it were, a scientist of Dr Alexander's standing surely knows the evidence that dreams and dream memories may be generated almost instantly at the point of awakening: they do not have to have occurred in real time. So on two counts it cannot be shown that he was having experiences while his neocortex was 'dead'. We need not accuse him of direct dishonesty over this; but he clearly isn't going out of his way to correct any misunderstanding which might arise and be helpful to his case.

At the same time he realises quite well at some level that in recounting his vision he isn't telling the strict and literal truth. Does he suppose that those high cheekbones possessed by his female guide were real cheekbones, of real bone with real marrow, fed by real blood with real haemoglobin oxygenated by contact with real air? No: if he thinks about it all (which you should discourage, by the way) he thinks the cheekbones were spiritual, or metaphorical; but he has to keep saying they were cheekbones, because if they weren't real, why should we suppose he was really in Heaven or that any of it was actual, rather than simply a set of absurd noodlings by a man who was half conscious? Once that gap has opened up between what a man says and what he knows to be the strict truth, the beetle of humbug is in place and we can begin to force the gap wider; keep him thinking that he has to make allowances for what simpler folk can cope with, and within a few years he'll be thinking of himself as peddling a mere metaphor for a sophisticated truth from which in fact all honesty and semblance of real religion has long since leached away. Any tiny, dangerous fragment of real mystical experience which Dr Alexander may have had will be quite lost under a heap of things he thinks it judicious or appropriate to say or think. It isn't likely, of course, that Dr Alexander had a genuine mystical experience of any kind, even in a fleeting, momentary form, but it's just the sort of cheating the Enemy indulges in now and then, letting the little vermin have an uncovenanted fragment of the truth in amongst the falsehoods which they actually sought out for themselves.

Of course you will want to keep Dr Alexander from realising quite how absurd his noodlings are; in this respect I think we can be proud of the work we have done within the education system. A man like Dr Alexander can pass right through to the highest level of academic attainment these days without ever reading Plato or encountering any of the classical texts or poetry which might enlarge his cramped imagination. If we were now to reproach such a man for expressing his vision in terms fit only for a child's comic paper, we need no longer fear that he will apologise and try harder, or even issue a manly rejection of the triviality of our complaint when set against the importance of his message; no, far likelier he will embark on a gratifyingly silly defence of comics as an imaginative medium and mature art form!

Naturally others, by contrast, can be encouraged to realise in full the absurdity of the noodlings: another great benefit of Dr Alexander's intervention is the heart it will put into our friends the atheists. If this is the tosh we are up against, they'll think, we need not work too hard. I dare say Professor Dawkins will allow himself an extra glass of port and perhaps decide he can cancel one or two of his more tedious public appearances. This is all good in two ways: first, of course we like the atheist cause to prosper; but also we like people to relax into their slogans and their easy rhetoric. We like the well-worn clichés. We don't want anyone impelled to try a bit harder to come up with new ideas on either side. The last thing we want is for people to be prompted into genuine fresh, open-ended thought; and I'm sure you'll take good care that never happens with Dr Alexander's revelations.

Your affectionate Uncle,

Screwtape

Apologies to C.S. Lewis, and to readers who notice that Screwtape's theological acuity and literary gifts seem to be a bit below his old level. I should add Lewis's customary warning to readers: Screwtape, like all devils, is an habitual liar and nothing he says should be taken at face value. In passing, it's worth noting another thing about near-death experiences: materialist champions of artificial intelligence are often accused of wanting to change the meaning of certain words - 'intelligence', 'learning' and so on - to favour themselves: but it seems that some on the other side are just as bad and want to monkey with the meaning of that simple, homely old word 'death'!

What about it, though: could our consciousness possibly survive death? There has always been rather strong evidence that consciousness arises from the activity of our brain. Stop the brain with a blow to the head and you get unconsciousness: stop it permanently with a slightly harder blow and you get death. Over the last hundred years or so the correlation has been worked out in more and more detail and the empirical evidence of exquisitely detailed correspondences is pretty overwhelming. Although we don't yet have a scientific explanation of consciousness, the size and shape of the gap where one might go is becoming clear. In fact, I should say it's now rather difficult to see how the kind of God envisaged by Lewis could be conscious; never having had to face the challenges of survival on the ancient savannah which presumably shaped the evolution of the human brain, and not having any neuronal apparatus, it's becoming hard to imagine how He (or Screwtape) could possibly manage it. As a matter of fact I suspect that Lewis and his fellow Inklings had a slightly unorthodox neoplatonic conception of God which they partly kept to themselves; but that's another subject.

We can arrive at a similar conclusion about our own physicality without being so brutally reductionist. When I sit down and think seriously about what day to day life as a disembodied spirit would be like, I find myself in some difficulty. With no eating, sleeping, walking around or working, I don't quite know what would be left of my life. Perhaps if there is a spiritual internet I could carry on blogging: surely the discussion of the mind itself is a sufficiently immaterial activity to go on after death? Well, maybe: but my motives for doing it all seem to be based in my animal nature. Curiosity - the way the primate brain goes on demanding to be fed, whereas the feline one obligingly goes to sleep once the stomach is full - is part of it; a vague fear that I don't basically know what is going on here (anywhere), which no doubt relates to simpler nervousnesses out on that ancestral savannah; the acquisitive pleasure of having 'got'

something, even if only an abstract argument; they all have their roots in biology and would probably cease to influence my ghost. If so, what would that ghost amount to, and how much would it have in common with the robustly biological animal which pressed the keys to create this text?

It could still be that although our mental life arises from neuronal activity it has a spiritual dimension, but it seems to me that the contest is almost a default win for materialism because traditional dualism barely offers any view as to how spiritual consciousness actually works. An honourable exception is the theory offered by Sir John Eccles with Karl Popper; but their psychons seem to me to mirror neurons a little too closely.

There's scope for a panpsychist theory whereby our consciousness relapses into its constituent parts after death; but although those parts would still be conscious to some degree it wouldn't be our distinctive individual consciousness, so to my taste it's not much better.

What then, about a thoroughly materialist survival? Could the details of our neuronal activity be recorded and then replayed in some suitable medium - or could our neurons gradually be replaced by silicon? I put aside here the Searlian theory that consciousness requires some as-yet unknown special property of neuronal tissue, because hypothetically we can grow artificial neuronal tissue and use that if it's really necessary. My concerns here are to do with identity. It might be possible to replay an exact copy, but a copy isn't me; and if I gradually get replaced, won't I be gradually phased out in a similar way? The fact that I might never know it and might be replaced by a very similar individual is not really enough. I put aside here also the idea that neuronal replacement is theoretically impossible, that the process of neuronal interaction is such that for esoteric reasons it cannot be stopped and restarted or bits replaced; although that's an interesting hypothesis it doesn't look as if things work like that.

It does look, though, as if our essential neurons never do get replaced in the normal course of events. We do actually correspond in relatively simple physical terms to a definite set of neurons that last us our whole life; and when they go, we go. Without proving anything absolutely, I think that gives quite a bit of weight to my fears about the effect of substitution on my identity.

It's not life after death, but what, for the sake of completeness, about just going on as we are: what about getting that same set of neurons to last forever? Let us, as Woody Allen put it, not live on through our work, but live on through not dying. I don't know why that should be impossible, but I suspect it is. It is noteworthy that remaining young and fit indefinitely should surely have the best possible survival value; yet all the large organisms produced by evolution die. It looks as if we're up against some fundamental constraint.

My own conclusion is that death is indeed inevitable, final, and definitive. That is an unpleasant prospect, but perhaps when I am grown up I will at last be calm and clear enough about it to be able to tell the assorted merchants of immortality what I think is actually the final truth: one good life is enough.

Blind Brain

12 November 2012

Besides being the author of thoughtful comments here – and sophisticated novels, including the great fantasy series The Second Apocalypse – Scott Bakker has developed a theory which may dispel important parts of the mystery surrounding consciousness.

This is the Blind Brain Theory (BBT). Very briefly, the theory rests on the observation that from the torrent of information processed by the brain, only a meagre trickle makes it through to consciousness; and crucially that includes information about the processing itself. We have virtually no idea of the massive and complex processes churning away in all the unconscious functions that really make things work and the result is that consciousness is not at all what it seems to be. In fact we must draw the interesting distinction between what consciousness is and what it seems to be.

There are of course some problems about measuring the information content of consciousness, and I think it remains quite open whether in the final analysis information is what it's all about. There's no doubt the mind imports information, transforms it, and emits it; but whether information processing is of the essence so far as consciousness is concerned is still not completely clear. Computers input and output electricity, after all, but if you tried to work out their essential nature by concentrating on the electrical angle you would be in trouble. But let's put that aside.

You might also at first blush want to argue that consciousness must be what it seems to be, or at any rate that the contents of consciousness must be what they seem to be: but that is really another argument. Whether or not certain kinds of conscious experience are inherently infallible (if it feels like a pain it is a pain), it's certainly true that consciousness may appear more comprehensive and truthful than it is.

There are in fact reasons to suspect that this is actually the case, and Scott mentions three in particular; the contingent and relatively short evolutionary history of consciousness, the complexity of the operations involved, and the fact that it is so closely bound to unconscious functions. None of these prove that consciousness must be systematically unreliable, of course. We might be inclined to point out that if consciousness has got us this far it can't be as wrong as all that. A general has only certain information about his army – he does not know the sizes of the boots worn by each of his cuirassiers, for example – but that's no disadvantage: by limiting his information to a good enough set of strategic data he is enabled to do a good job, and perhaps that's what consciousness is like.

But we also need to take account of the recursively self-referential nature of consciousness. Scott takes the view (others have taken a similar line), that consciousness is the product of a special kind of recursion which allows the brain to take into account its own operations and contents as well as the external world. Instead of simply providing an output action for a given stimulus, it can throw its own responses into the mix and generate output actions which are

more complex, more detached, and in terms of survival, more effective. Ultimately only recursively integrated information reaches consciousness.

The limits to that information are expressed as information horizons or strangely invisible boundaries; like the edge of the visual field the contents of conscious awareness have asymptotic limits – borders with only one side. The information always appears to be complete even though it may be radically impoverished in fact. This has various consequences, one of which is that because we can't see the gaps, the various sensory domains appear spuriously united.

This is interesting, but I have some worries about it. The edge of the visual field is certainly phenomenologically interesting, but introspectively I don't think the same kind of limit seems to come up with other senses. Vision is a special case: it has an orderly array of positions built in, so at some point the field has to stop arbitrarily; with sound the fading of farther sounds corresponds to distance in a way which seems merely natural; with smell position hardly comes into it and with touch the built-in physical limits mean the issue of an information horizon doesn't seem to arise. For consciousness itself spatial position seems to me at least to be irrelevant or inapplicable so that the idea of a boundary doesn't make sense. It's not that I can't see the boundary or that my consciousness seems illimitable, more that the concept is radically inapplicable, perhaps even metaphorically. Scott would probably say that's exactly how it is bound to seem…

There are several consequences of our being marooned in an encapsulated informatic island whose impoverishment is invisible to us: I mentioned unity, and the powerful senses of a 'now' and of personal identity are other examples which Scott covers in more detail. It's clear that a sense of agency and will could also be derived on this basis and the proposition that it is our built-in limitations that give rise to these powerfully persuasive but fundamentally illusory impressions makes a good deal of sense.

More worryingly Scott proceeds to suggest that logic and even intentionality – aboutness – are affected by a similar kind of magic that similarly turns out to be mere conjuring. Again, results generated by systems we have no direct access to, produce results which consciousness complacently but quite wrongly attributes to itself and is thereby deluded as to their reliability. It's not exactly that they don't work (we could again make the argument that we don't seem to be dead yet, so something must be working) more that our understanding of how or why they work is systematically flawed and in fact as we conceive of them they are properly just illusions.

Most of us will, I think want to stop the bus and get off at this point. What about logic, to begin with? Well, there's logic and logic. There is indeed the unconscious kind we use to solve certain problems and which certainly is flawed and fallible; we know many examples where ordinary reasoning typically goes wrong in peculiar ways. But then there's formal explicit logic, which we learn laboriously, which we use to validate or invalidate the other kind and which surely happens in consciousness (if it doesn't then really I don't think anything does and the whole matter descends into complete obscurity); hard not to feel that we can see and understand how that works too clearly for it to be a misty illusion of competence.

What about intentionality? Well, for one thing to dispel intentionality is to cut off the branch on which you're sitting: if there's no intentionality then nothing is about anything and your theory has no meaning. There are some limits to how radically sceptical we can be. Less fundamentally, intentionality doesn't seem to me to fit the pattern either; it's true that in

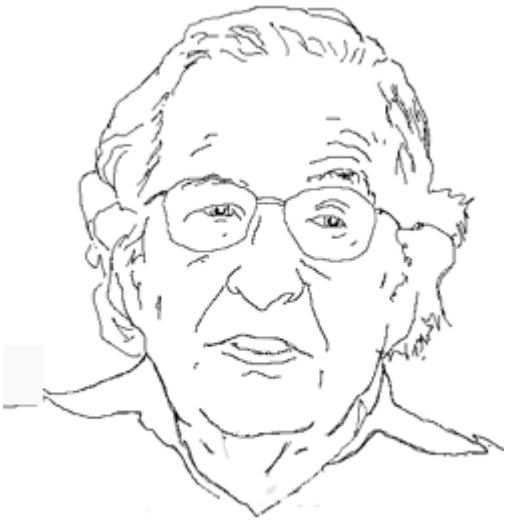
everyday use we take it for granted, but once we do start to examine it the mystery is all too apparent. According to the theory it should look as if it made sense, but on the contrary the fact that it is mysterious and we have no idea how it works is all too clear once we actually consider it. It's as though the BBT is answering the wrong question here; it wants to explain why intentionality looks natural while actually being esoteric; what we really want to know is how the hell that esoteric stuff can possibly work.

There's some subtle and surprising argumentation going on here and throughout which I cannot do proper justice to in a brief sketch, and I must admit there are parts of the case I may not yet have grasped correctly – no doubt through density (mine, not the exposition's) but also I think perhaps because some of the latter conclusions here are so severely uncongenial. Even if meaning isn't what I take it to be, I think my faulty version is going to have to do until something better comes along.

(BTW, the picture is supposed to be Thomas Aquinas, who introduced the concept of intentionality. The glasses are suppose to imply he's blind, but somehow he's just come out looking like a sort of cool monk dude. Sorry about that.)

Chomsky on AI

25 November 2012



There was an interesting conversation with Noam Chomsky published recently. The introductory piece mentions the review by Chomsky which is often regarded as having dealt the death-blow to behaviourism, and leaves us with the implication that Chomsky has dominated thinking about AI ever since. That's overstating the case a bit, though it's true the prevailing outlook has been mainly congenial to those with Chomskian views . What's generally taken to have happened is that behaviourism was succeeded by functionalism, the view that mental states arise from the functioning of a system - most often seen as a computational system. Functionalism has taken a few hits since then, and a few rival theories have emerged, but in essence I think it's still the predominant paradigm, the idea you have to address one way or another if you're putting forward a view about consciousness. I suspect in fact that the old days, in which one dominant psychological school - associationism, introspectionism, behaviourism - ruled the roost more or less totally until overturned and replaced equally completely in a revolution, are over, and that we now live in a more complex and ambivalent world.

Be that as it may, it seems the old warrior has taken up arms again to vanquish a resurgence of behaviourism, or at any rate of ideas from the same school: statistical methods, notably those employed by Google. The article links to a rebuttal last year by Peter Norvig of Chomsky's criticisms, which we talked about at the time. At first glance I would have said that this is all a non-issue, because nobody at Google is trying to bring back behaviourism. Behaviourism was explicitly a theory about human mentality (or the lack of it); Google Translate was never meant to emulate the human brain or tell us anything about how human cognition works. It was just meant to be useful software. That difference of aim may perhaps tell us something about the way AI has tended to go in recent years, which is sort of recognised in Chomsky's suggestion that it's mere engineering, not proper science. Norvig's response then was reasonable but in a way it partly validated Chomsky's criticism by taking it head-on, claiming serious scientific merit for 'engineering' projects and for statistical techniques.

In the interview, Chomsky again attacks statistical approaches. Actually 'attack' is a bit strong: he actually says yes, you can legitimately apply statistical techniques if you like, and you'll get results of some kind - but they'll generally be somewhere between not very interesting and meaningless. Really, he says, it's like pointing a camera out of the window and then using the pictures to make predictions about what the view will be like next week: you might get some good predictions, you might do a lot better than trying to predict the scene by using pure physics, but you won't really have any understanding of anything and it won't really be good science. In the same way it's no good collecting linguistic inputs and outputs and matching everything up (which does sound a bit behaviouristic, actually), and equally it's no good drawing statistical inferences about the firing of millions of neurons. What you need to do is find the right level of interpretation, where you can identify the functional bits - the computational units - and

work out the algorithms they're running. Until you do that, you're wasting your time. I think what this comes down to is that although Chomsky speaks slightingly of its forward version, reverse engineering is pretty much what he's calling for.

This is, it seems to me, exactly right and entirely wrong in different ways at the same time. It's right, first of all, that we should be looking to understand the actual principles, the mechanisms of cognition, and that statistical analysis is probably never going to be more than suggestive in that respect. It's right that we should be looking carefully for the right level of description on which to tackle the problem - although that's easier said than done. Not least, it's right that we shouldn't despair of our ability to reverse engineer the mind.

But looking for the equivalent of parts of a Turing machine? It seems pretty clear that if those were recognisable we should have hit on them by now, and that in fact they're not there in any readily recognisable form. It's still an open question, I think, as to whether in the end the brain is basically computational, functionalist but in some way that's at least partly non-computational, or non-functionalist in some radical sense; but we do know that discrete formal processes sealed off in the head are not really up to the job.

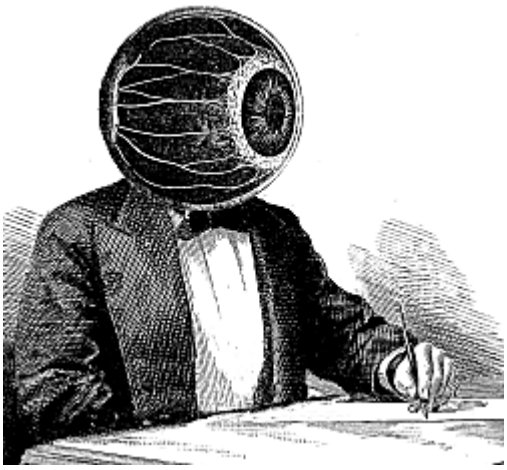
I would say this has proved true even of Chomsky's own theories of language acquisition. Chomsky, famously, noted that the sample of language that children are exposed to simply does not provide enough data for them to be able to work out the syntactic principles of the language spoken around them as quickly as they do (I wonder if he relied on a statistical analysis, btw?). They must, therefore, be born with some built-in expectations about the structure of any language, and a language acquisition module which picks out which of the limited set of options has actually been implemented in their native tongue.

But this tends to make language very much a matter of encoding and decoding within a formal system, and the critiques offered by John Macnamara and Margaret Donaldson (in fact I believe Vygotsky had some similar insights even pre-Chomsky) make a persuasive case that it isn't really like that. Whereas in Chomsky the child decodes the words in order to pick out the meaning, it often seems in fact to be the other way round; understanding the meaning from context and empathy allows the child to word out the proper decoding. Syntactic competence is probably not formalised and boxed off from general comprehension after all: and chances are, the basic functions of consciousness are equally messy and equally integrated with the perception of context and intention.

You could hardly call Chomsky an optimist: It's worth remembering that with regard to cognitive science, we're kind of pre-Galilean, he says; but in some respects his apparently unreconstructed computationalism is curiously upbeat and even encouraging.

Spaun

9 December 2012



The first working brain simulation? Spaun (Semantic Pointer Architecture Unified Network) has attracted a good deal of interested coverage.

Spaun is based on the Nengo neural simulator: it basically consists of an eye and a hand: the eye is presented with a series of pixelated images of numbers (on a 28 x 28 grid) and the hand provides output by actually drawing its responses. With this simple set up Spaun is able to perform eight different tasks ranging from copying the digit displayed to providing the correct continuation of a number sequence in the manner of certain IQ tests. Its performance within this limited repertoire is quite impressive and the fact that it fails in a few cases actually makes it resemble a human brain even more closely. It cannot learn new tasks on its own, but it can switch between the eight at any time without impairing its performance.

Spaun seems to me an odd mixture of approaches; in some respects it is a biologically realistic simulation, in others its structure has just been designed to work. It runs on 2.5 million simulated neurons, far fewer than those used by purer simulations like Blue Brain; the neurons are essentially designed to work in a realistic way, although they are relatively standardised and stereotyped compared to their real biological counterparts. Rather than being a simple mass of neurons or a copy of actual brain structures they are organised into an architecture of modules set up to perform discrete tasks and supply working memory, etc. If you wanted to be critical you could say that this mixing of simulation and design makes the thing a bit kludgeish, but commentators have generally (and rightly, I think) not worried about that too much. It does seem plausible that Spaun is both sufficiently realistic and sufficiently effective for us to conclude it is really demonstrating in practice the principles of how neural tissue supports cognition - even if a few of the details are not quite right.

Interesting in this respect is the use of semantic pointers. Apparently these are compressions of multidimensional vectors expressed by spiking neurons; it looks as though they may provide a crucial bridge between the neuronal and the usefully functional, and they are the subject of a forthcoming book, which should be interesting.

What's the significance of Spaun for consciousness? Well, for one thing it makes a significant contribution to the perennial debate on whether or not the brain is computational. There is a range of possible answers which go something like the following.

1. Yes, absolutely. The physical differences between a brain and a PC are not ultimately important; when we have identified the right approach, we'll be able to see that the basic activity of the brain is absolutely standard Turing-style computations.
2. Yes, sort of. The brain isn't doing computation in quite the way silicon chips do it, but the functions are basically the same, just as a plane doesn't have flapping feathery wings but is still doing the same thing - flying - as a bird.

3. No, but. What the brain does is something distinctively different from computation, but it can be simulated or underpinned by computational systems in a way that will work fine.
4. No, the brain isn't doing computations and what it is doing crucially requires some kind of hardware which isn't a computer at all, whether it's some quantum gizmo or something with some other as yet unidentified property which biological neurons have.

The success of Spaun seems to me to lend a lot of new support to position 3: to produce the kind of cognitive activity which gives rise to consciousness you have to reproduce the distinctive activity of neurons - but if you simulate that well enough by computational means, there's no reason why a sufficiently powerful computer couldn't support consciousness.

Pockett Redux

16 December 2012

Sue Pockett is back in the JCS with a new paper about her theory that conscious experience is identical with certain electromagnetic patterns generated by the brain. The main aim of the current paper is to offer a new hypothesis about what distinguishes conscious electromagnetic fields from non-conscious ones; basically it's a certain layered shape, with negative charge on top followed by a positive region, a neutral one, and another positive layer.

Pockett sets the scene by suggesting that there are three main contenders when it comes to explaining consciousness: the thesis that consciousness is identical with certain patterns of neuron firing; functionalism, the view that consciousness is a process; and her own electromagnetic field theory. This doesn't seem a very satisfactory framing. For one thing it doesn't do much justice to the wild and mixed-up jungle which the field really is (or so it seems to me); second there isn't really a sharp separation between the first two views she mentions - plenty of people who think consciousness is identical with patterns of neuronal activity would be happy to accept that it's also a neuronal process. Third, it does rather elevate Pockett herself from the margins to the status of a major contender; forgiveable self-assertion, perhaps, although it may seem a little ungenerous that she doesn't mention that others have also suggested that consciousness is an electromagnetic field (notably Johnjoe McFadden - although he believes the field is causally effective in brain processes, whereas Pockett, as we shall see, regards it as an epiphenomenon with no significant effects).

Pockett mentions some of the objections to her theory that have come up. One of the more serious ones is the point that the fields she is talking about are in fact tiny and very local: so local that there is no realistic possibility of the field from one neuron influencing the activity of another neuron. Pockett is happy to accept this; conscious experience doesn't actually have the causal role we usually attribute to it, she says: it's really just a kind of passenger accompanying mental processes whose course is determined elsewhere. She cites various people in support of this epiphenomenalist view, including Libet - she has some interesting things to say about his experiments (but doesn't note that he too liked the idea of a mental field of some kind).

The new thesis about the shape of conscious fields appears to spring from observation of neuronal structure in the neocortex, and in particular the fact that in its fourth layer there are no pyramidal cell bodies. This is a feature which appears to be characteristic of the parts of the neocortex often associated with consciousness, and it implies that there is a gap or neutral region in the fields generated there, helping to yield the layered structure mentioned above. Pockett proposes a number of experiments which might support her view, some of which have already been carried out.

So what do we make of that? I see a number of problems.

First is the inherent implausibility of the overall theory. Why on earth would we identify conscious experience with tiny electromagnetic fields? I think the attraction of the theory

comes from thinking about two kinds of consciousness, more or less the two kinds identified by Ned Block: the a-consciousness that does the work of useful, practical cogitation, and the p-consciousness which simply endows it with subjective feeling. Looked at from this angle it may be appealing to think that the actual firing of neurons is doing the a-consciousness while the subjectivity, the phenomenal experience, come from the subsidiary electric buzz.

But when we look at it more closely, that doesn't make any particular sense. The problem is that conscious experience doesn't seem like anything in physics, whether a field or a lump of matter. If we're going to identify it with something physical, we might as well identify it with the neurons and simplify our theory - dropping the field entity in obedience to Occam's Razor. Nothing about the electrical fields helps to explain the nature of phenomenal experience in a way that would motivate us to adopt them.

Second, there's the problem of Pockett's epiphenomenalism. Epiphenomenalism is very problematic because if consciousness does not cause any of our actions, our reports and discussions about it cannot have been caused by it either. Pockett's own account of consciousness could not have been caused by her actual conscious experience, and nothing she writes could be causally related to that actual experience: if she were a zombie with no consciousness, she would have written just the same words. This is a bit of an uncomfortable position in general, but it also means that Pockett's ambitions to test her theory experimentally are doomed. You cannot test scientifically for the existence of a supposed entity that has no effects because it makes no observable difference whether it's there or not. All of Pockett's proposed experiments, on examination, test accessible aspects of her theory to do with how the fields are generated by neurons: none of them test whether consciousness is present or absent, and on her theory no such experiment is possible.

While those issues don't technically prove that Pockett is wrong, they do seem to me to be bad problems that make her theory pretty unattractive.