Conscious Entities

The Archive

2007 to 2009

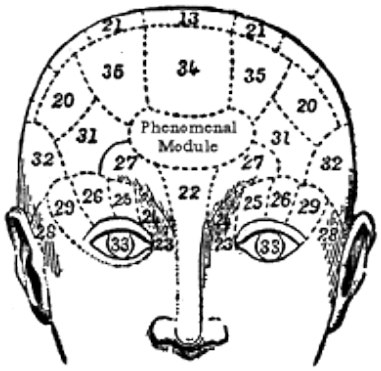
Welcome to Conscious Entities, the Archive

This is the second 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 A few posts have been omitted, and quite a few are rather out of date, but I still enjoy reading them.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.

The Genuine Problem

4 January 2007



Pete Mandik has posted a thought-provoking paper by Jack, Robbins, and Roepstorff which suggests we may have been considering the wrong issue all along. The problem with the Hard Problem (how do we square our ineffable, subjective experience of the world with the mechanical reality described by physics), they say, is that we tend to regard subjective experiences as being out there in the same sort of way as physical objects. This makes it hard for us to understand how our two pictures of the world can be reconciled. We end up looking for a mysterious missing ingredient in subjective experience, but that search is hopeless. In fact, JR&R suggest, the difference between the two accounts of the world arises from our using two different brain modules: one aimed at the world in general, one aimed specifically at phenomenal states.

That seems plausible enough at first sight and JR&R contend that it is a parsimonious theory too. It does require an additional brain module, but if you assume that the alternative is some form of dualism (as I think they do) then they're right, since the additional ontological commitment invoved in dualism would easily outweigh the merely neurological one required for an extra brain module. Moreover, there is apparently some good evidence to support the existence of the phenomenal brain module. It has been shown that activity in parts of the brain concerned with the external world correlates negatively with activity in the parts concerned with thinking about our own mental states (not too surprising, this - it's hard to imagine paying close attention to your own feelings and to the details of what is going on around you at the same time). More dubiously, JR&R suggest that autism looks a bit like what you get when your phenomenal module fails to operate correctly.

This doesn't seem quite right, however. If your phenomenal module ceases to function, you surely ought to become a philosophical 'zombie' - someone who has no subjective experience. That wouldn't be at all like autism, however. The behaviour of a philosophical zombie is perfectly normal (since your behaviour is determined by your non-phenomenal cognition): autism, however, certainly does affect your behaviour, in some cases very severely.

The problem is that JR&R are actually assigning three distinct roles to their module: they want it to provide phenomenal experience, to be a kind of higher-order facility which tells us about our own mental states, and a theory-of-mind machine which enables us to understand other people and social interactions (the bit most relevant to autism). The paper, I think, is a little light on explaining why these three things arise from the same basic function - in fact it almost seems to treat them as evidently equivalent. In fairness the paper doesn't pretend to be more than a sketch of quite a wide-ranging set of ideas.

Do the three go together? I suppose the insight that links them all is that knowing how something feels to us helps us understand how similar experiences feel to other people (only helps, though - I think our understanding of other people consists of a good deal more than just empathy). It is certainly plausible that our understanding of our own mental states arises from our understanding of other people's (though there are those who would say that it is our understanding of other people's minds that leads us to think we have our own). Less persuasive

on the face of it is the view that our subjective experience is a matter of knowledge about our own inner states. My subjective experiences appear to me to be about the external world for the most part, and it isn't immediately clear why second-order knowledge of my own mental states should endow them with subjective qualities. Of course, some people have put forward theories very much along those lines - Nicholas Humphrey, for example. But you certainly can't, as it were, have that conclusion for nothing.

Anyway if JR&R are at least broadly right, then there will always appear to be a mysterious Hard Problem, because we're just built that way. But they hold out instead the possibility of addressing instead the 'Genuine Problem', namely the question of the structure of cognition and its two modules. The good news, they say, is that this question can be addressed scientifically, so we won't have to wait around to see whether philosophers can get anywhere with the issues over the next thousand years or so. As a project, this is unquestionably a good idea: if science could explain the differences between the two modules and how one gives rise to subjectivity, that would be a very major advance. Unfortunately, I think merely saying that makes it clear how much remains to be done, and raises a fear that JR&R have themselves fallen into a trap they describe: of setting out to explain qualia and ending up explaining something more amenable to science instead.

JR&R also make a plea for the return of the subjective as a field of proper research, mentioning the introspectionists of bygone days. Rhetorically this may be a mistake: I found my own automatic reaction was more or less the same as if they had called for a fresh look at the virtues of Ptolemaic astronomy. In fact they are careful to distinguish between the problematic efforts of Titchener and Wundt and the more measured approach they advocate.

A stimulating paper, anyway, though I for one will continue to beat my head philosophically against the good old Hard Problem.

Mind the Gap

18 January 3007

The Royal Society had a joint event with the Royal Society of Literature back in November, devoted to the question of the mental 'gap' between human beings and animals. An eminent panel (Nicky Clayton, Doris Lessing, Will Self and Andrew Whiten) offered thoughts and addressed a few questions.

In fact, although entertaining and somewhat enlightening, the discussion never made more than glancing contact with any of the philosophical issues; a lot of the time was spent in remarking on the charm and intelligence of apes, corvids, cats, and other animals. Will Self asserted that there were no differences between apes and humans. Having read his novel Great Apes, which provides a number of interesting details about the social and sexual behaviour of chimpanzees, I thought this shed a surprising light on his own private life; as did, perhaps, his insistence that consciousness was greatly over-rated.

Anyway, I thought it might be worth trying to scratch the surface of the issues a little more deeply. What are the fundamental mental differences between animals and humans? For our purposes three areas stand out: moral, intellectual, and phenomenal.

¶A number of moves to upgrade the official moral status of animals have been debated in recent times: the Great Ape Project, for example, seeks to obtain recognition of three basic legal rights for our nearer relatives among the animals. Whether this is philosophically a sound approach is open to doubt; Roger Scruton, at least, has argued that since animals do not have moral duties, they cannot have rights either. I find the idea that duties and rights go together in this way very attractive: unfortunately, I can't see any compelling reason to think it's true. However, it could reasonably be argued that for talk about rights to be meaningful, the person with those rights has to have the ability to assert or give up the right in question, and some animals surely don't have the intellectual capacity to assert or give up anything that abstract. How far this is true of all animals, of course, depends on your view of their intellectual capacities, and to some degree on how explicit you feel the assertion of rights needs to be - is struggling to survive an assertion of the right to live?

We don't have to talk in terms of rights, however, and in fact the idea of natural or moral rights is a slippery concept which sometimes gives a false air of dignity to mere desires or opinions - when I say animals have a right to something, I may be doing no more than urging others to agree to giving it. Peter Singer's line on animal morality takes a different tack, making them parties to Utilitarianism. The argument here is that animal pain or pleasure should be weighed in the scales together with human feelings. If you're a utilitarian, this conclusion seems a logical extension of the system, and even if you're not you may well feel that animal pain is morally relevant. However, the argument does rest on the assumption that animals feel pain, which has been denied by some, about some animals in any case; so this line of argument may rest on what view we take of the phenomenal capacities of animals and how far their version of pain resembles ours (although it would be perfectly possible, logically, to adopt a form of utilitarianism which took no account of phenomenal qualities and merely took it that pain responses had negative utility regardless of whether they 'really hurt').

One argument which tends to lend weight to a Scrutonish line of thinking is the plausible contention that animals are moral objects, but not moral subjects: that is, there may be moral rules about what can be done to them, but not about what they should do. It may be a moral principle that a wolf should not be caused unnecessary pain; but we don't consider holding wolves accountable for their behaviour and subjecting them to the apparatus of the criminal justice system for any excess pain they may have caused to other animals.

Or do we? In training and working with a dog we certainly do use punishments and rewards: we talk in censorious tones of bad behaviour and admiring ones of good: not all that different from the way we might behave towards a child. There's scope for argument here over whether we really think a 'bad dog' is morally bad or bad in some lesser, more pragmatic sense - whether, to get a bit Kantian, we're dealing with categorical or merely hypothetical imperatives.

It seems reasonable to me, skipping over that discussion, that animals fall on a moral scale so far as responsibility is concerned: most animals have a negligible level of moral responsibility, but the most intelligent show at least the beginnings of a capacity for it, albeit at a level well below that of most adult human beings.

We seem, anyway, to have established a pleasingly symmetrical (but suspiciously neat) structure whereby there are two lines to take on the moral qualities of animals, each of which depends at least partly on the view we take of one of the other two areas I mentioned to begin with.

So what about the intellectual capacities of animals? The Royal Society discussion touched on the way tool use, once considered a uniquely human practice, has gradually been found to occur in a range of species and a number of different forms (Including, apparently, the nose-picking tool, something not yet achieved by human technology so far as I know. I hope the designers of the Swiss Army Knife are not reading this.) This is undoubtedly true, but it still does seem that human tool use is distinctively different, even if it's difficult to formulate the distinction in qualitative terms.

The subject of animal linguistic competence causes strong disagreement. The work of Sue Savage-Rumbaugh and others has produced chimps which appear to manage a high degree of communication through special keyboards or through sign language; but while some find the results compelling evidence of real linguistic competence verging on the human, others remain unconvinced.

It strikes me that there is an interesting comparison to be made here with the Turing test. Both candidates for human-style cognition, the chimp and the computer, are to be judged very largely on their ability to hold a conversation (maybe we're not so much Homo Sapiens as Homo Loquens), and in both cases the results are interpreted very differently according to the predisposition of the observer. However, it seems to me we expect much more grammatical competence from a computer than we do from a chimpanzee: the Turing test requires properly constructed human sentences, whereas the chimpanzees are provided with a simplified system where it's much harder to go wrong. In fact, we require the machine's text responses to be indistinguishable from those of a human being, but a few errors on the chimp's part would not cause us to rise from the desk, satisfied that the creature was a sham.

In a different respect, though, we're harder on the primate: we take it for granted that if the chimp can communicate at all it will be able to tell us about objects nearby and carry out simple

instructions even in unfamiliar surroundings: to ask the computer to fetch a ball and give it to its sister would not be considered fair play.

I think the comparison of our attitudes in the two cases suggests that we do, almost unthinkingly, attribute to apes some of the basic mental abilities we enjoy ourselves: but the fact that we let them off the tougher linguistic tasks we expect an inanimate algorithm to be able to deal with shows that even the pro-primate school of thought is aware of a distinct gap.

What then, about the phenomenal experience of animals? There is a pretty well established consensus for practical purposes which says the bigger your brain, the more you feel and the greater the care that must be taken to avoid unnecessary suffering. This may short-change some non-mammals a bit, but it seems an appealing rule of thumb. A minority school of thought says that humans are qualitatively different here; only humans know that they are in pain, and that's what makes it so bad. Dennett, who as an enemy of qualia in all their forms, wants pain to be a matter of interpretation, seems to say it's largely a matter of crushed hopes and projects, though I don't think he goes on to draw the implied negative conclusion about the pain capacity of projectless animals.

The real philosophical problem is that we have no secure grounds for believing anyone feels pain except ourselves: it's only a weak induction from the single case where we have direct evidence. Perhaps the best we can say is that the induction applies to some animals almost as well as it applies to our fellow human beings.

Where does that leave us? Are animals conscious? I don't know, but my guess is that there is no clear qualitative gap between humans and animals - there's nothing we've got that at least some of them haven't got in at least a vestigial form. But there's a huge quantitative gap; they're never more than slightly conscious.

All Done With Mirrors

January 29 2007

Vilayanur S. Ramachandran has a short piece on Edge about the neurology of self-awareness - more specifically about the role played by mirror neurons. Ramachandran seems to have a special interest in mirrors, having used them in a famous series of experiments which succeeded in eliminating the pain from the 'phantom limbs' sometimes experienced by amputees. He certainly attaches considerable importance to mirror neurons.

We've known for some time the relatively unsurprising fact that certain groups of neurons fire whenever an experimental subject performs a certain action. More recently it has emerged that some of these neurons also fire when the subject sees someone else performing the same action. It's as though when the brain sees an action performed, it goes through a small pantomime of triggering the same action; it mimics or mirrors the presumed brain activity of the person being observed. Similarly, among the neurons that fire when a subject is poked, there are some which also fire when the subject sees someone else being poked.

These mirror neurons are clearly interesting in a number of ways, but perhaps the most striking is that they appear to provide a clear neurological basis for empathy, and perhaps for the ability humans (and a few other animals) have to reason successfully about other points of view. This capacity is often called a 'theory of mind', or 'theory of other minds' though I think that's a misleading label which implies a much more explicit understanding than is actually at work. Being able to tell what your rivals know and don't know, and how they are likely to behave as a result, clearly opens up a whole new field of opportunities for a cunning organism, but it involves a level of abstraction which few animals appear to have reached, and indeed it seems the ability does not become fully developed even in humans until they are well into early childhood.

The mirror neurons we know about to date are not, of course, enough by themselves to provide any very high-flown level of empathy, but they suggest that similar processes may be at work on a higher level, and they might well be part of the final answer. Ramachandran wants to go a bit further, however, and suggest that they help constitute our sense of self. I think it would be true to say that the traditional view here is that we begin as solipsists, with a natural sense of self but not really distinguishing other people from the inanimate objects around us. Then we go on to make the awesome conceptual leap into realising that there are other people like ourselves out there, and that they have thoughts and feelings similar to our own (with some psychopaths, perhaps, never making the leap and remaining permanently indifferent to other people's pain).

Ramachandran's proposal is that the process works the other way: we start by observing the behaviour of other people and come to have a basic understanding of them sufficient to attribute a kind of selfhood to each of them. It's only then that the empathy supported by mirror neurons leads us to realise that we too have a self of the same kind. Ramachandran

acknowledges that others have offered theories along roughly similar lines, but I think this is the first time the link with mirror neurons has been drawn in this way.

Are these ideas right? I think it's important to be clear about what kind of selfhood we're talking about here. In spite of the poking experiments, I don't think it can be anything to do with our experiences, or feelings. Take pain: if we applied the theory here we should be saying that to begin with we may be aware of pain, but don't really attribute it to ourselves: noticing how other people try to avoid it at all costs, we begin to think that the pian we experience actually belongs to us. I suppose it could be so, but knowing as we do what pain is actually like, and how one of its properties is a location in our body, it seems implausible to me. That in turn casts some doubt on whether this kind of explanation from others to ourselves can deal with important cases like our ability to tell that what other people can see is different from what we ourselves are aware of. It doesn't seem very likely, in other words, that we come to know there are things hidden from our view because we have previously noticed that things may be hidden from other people.

I could be wrong about that, but perhaps we are really talking about the self as the origin of volition; the mysterious thing that makes the decisions and (at least when things are running normally) does the talking. We attribute the intentions and thoughts of other people to an essential core called the self, and through empathy we come to attribute our own to a similar self. This seems a much more appealing line of argument.

There is certainly something hard to pin down about our selfhood in this sense. Hume famously observed that when he tried to observe his inner self there didn't seem to be anything there except a bundle of perceptions; Dennett has suggested that the self is a kind of explanatory abstraction akin to a centre of gravity. I think the problem is that the self in this sense is not so much a thing as a source - like the spring which provides the origin of a river. Hume says all he can see is a lot of water; Dennett says the source is a geometrical abstraction, not a real physical thing: there's something in both points of view but both also have an element of perversity.

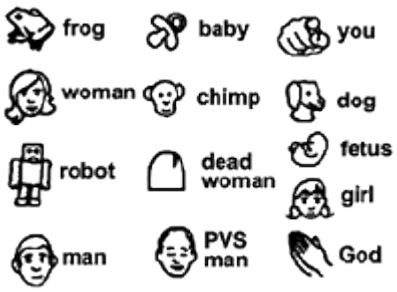
I'll make the bold claim here that Hume actually missed something, in that sometimes we actually experience the emergence of thoughts and intentions. I submit that we sometimes know, in an inexplicit way, what we are about to say, and even what we are about to think about, before the event actually occurs; we directly experience the emergence of intentional stuff, and hence have a direct handle on our own selfhood. I must admit that this claim, resting as it does on introspective evidence, is not very strongly supported, but if it is true, it contradicts Ramachandran's view: we know about our selfhood from direct experience, not from observing others, and in a way quite different from the way we know about the selfhood of others.

It may be that Ramachandran is in fact thinking about yet another kind of selfhood; apart from the two I have touched on. But there is a further reason for doubting whether mirror neurons are as important as they seem. A lot depends here on the way the story unfolds. First we have neurons that fire as part of the neurological business of performing a task, or of experiencing a sensation. Then it turns out some of them fire when we merely see someone else performing the action, or having the experience. We draw the conclusion that these neurons are reflecting, or simulating, what's going on in the other person's brain; doing a subliminal imitation of what the other person is going through. But perhaps we ought to reinterpret our original view that the firing of these neurons is part of the business of performing the action. Perhaps these neurons were only ever part of a system for recognising actions. There was a perfect correlation between

them firing and our performing the action, but that was because every time we did the action, we recognised it: the neurons were never part of the actual performance of the action. It's then not surprising that the same neurons fire when we see the action performed by someone else. It remains interesting that a single set of neurons respond to the same action whether we are acting, or someone else; but once we shed the preconception that these neurons are part of our own control system, some of the wider implications fall away.

Dimensions of Mind

10 February 2007



A short paper by Gray, Gray and Wegner in Science recently sets out the results of an interesting survey of how people view minds. People were asked to make comparisons amongst a strange group of 13 miscellaneous entities (see picture), rating them against a series of criteria.

The main finding is that people seem to rank minds along two scales rather than one. Analysis of the data suggested that the two basic qualities of minds as perceived by the respondents were: experience (hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, joy) and agency (self-control, morality, memory, emotion recognition, planning, communication, and thought). This result is agreeably in line with philosophical thinking about consciousness; I don't think it would be too much of a stretch to claim that the 'hard' and 'easy' problems of consciousness are the problems of experience and agency respectively.

That's nice, but having one's preconceptions confirmed so neatly might be grounds for suspicion. Is it possible, for example, that the authors built in the distinction which emerged from the research by their choice of examples? The list of entities for the original experiment is certainly a strange one (in some respects the online version is even stranger: it consists mainly of living human beings, but also features two birds - surely unlikely to score very differently). Clearly these are meant to be a set of examples of special interest from a mental point of view, but the list is certainly not exhaustive: we could add aliens, ghosts, computers and a number of others. In fairness I must admit I don't know exactly what an exhaustive list would be in this context: but to illustrate the possibility of skewing the results, consider what might happen if we added the following (ghost, djinn, angel, hallucination, my mirror image, chat-bot program) it seems possible that something like solidity might have emerged as another apparently salient quality of minds, which is clearly incorrect (or is it?).

It's difficult, moreover, with this kind of research, to be sure extraneous considerations are being kept out. Take the results about God. God rates high on agency but very low on experience. I think this must be because negative items like pain and hunger feature prominently, and people found the idea of God experiencing pain unlikely. It might be that they thought of pain as the province of creatures with bodies, but I suspect that to some extent they were just distracted by the observation that an omnipotent, omniscient being ought to be well able to keep out of the way of pain - which is beside the point, strictly speaking. It's an odd result to get from a largely Christian group, in any case, given that Jesus surely experienced pain and hunger, if perhaps not rage

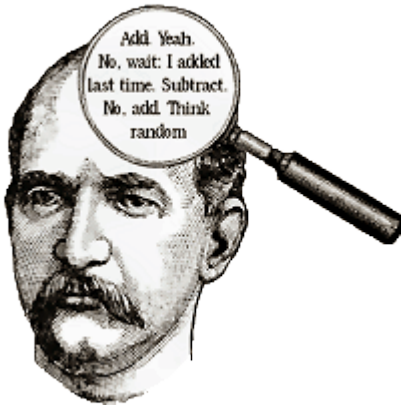
In fact, on this showing God is more or less a philosophical zombie: all agency and no experience. This raises another issue: if you have a problem with the concept of such zombies, you might be inclined to deny that the two dimensions identified here are really independent, arguing that agency actually implies experience. Presumably you might also be sceptical about the possibility of anti-zombies - creatures with full experience but no agency. I think, though, that I'd be I'd be inclined to reverse the sceptical argument and see the research as providing some good evidence against the assertion that zombies are inconceivable. They may well, on further consideration, and given some further argumentation, turn out to be impossible (as a

matter of fact I think they do); but this research makes it difficult to argue that they are outright incoherent or unimaginable. It's not very often that you get empirical results which bear on matters of philosophical interest, so this is surely some cause for celebration.

Quibbles aside, moreover, I think the results are essentially correct. I'm wondering now if an attempt to construct an 'exhaustive' list of interestingly different examples of consciousness/non-consciousness might itself be a useful, if perhaps doomed, exercise.

Decoding Intentions

20 February 2007



There was a good deal of media attention recently for Reading Hidden Intentions in the Human Brain (Haynes et al, Current Biology 2007), a paper which was said to herald the possibility of mind-reading. My initial reaction, I must admit, was sceptical from two directions. First, this seemed like another example of the hyperbole which afflicts the field (I particularly remember British Telecom announcing the possibility of a 'soul catcher' which would would record the electrical activity of the brain and hence - of course - capture one's very identity. I don't think they had a neurologist on the team.) Second, there's actually nothing new about being able to detect neural precursors of a physical action before it occurs.

However, on examination the research is both new and interesting, if a little less earth-shattering than reports might have suggested. It is based on MRI scanning but uses new techniques to extract a finer level of detail from the same hardware; and it detects an abstract intention rather than the precursors of a physical movement.

The experiments went like this: the subjects were told they would be given two numbers; they were to choose now whether they would add them or subtract the smaller from the larger. They then had to keep this intention in mind during a variable interval before the actual numbers were displayed, and in a final stage indicate from a choice of four numbers (one the correct result of addition, one of subtraction, and two irrelevant numbers) the result of their chosen calculation (from which, as the paper says, their choice of addition or subtraction could reasonably be inferred). Applying pattern recognition to the scans, the researchers found that they were able to identify the subjects' choices with up to 70% accuracy. They were, incidentally, able to do the same from a slightly different region of the brain while the subjects were indicating the results; but the interesting part is the ability to detect a difference between the two options before the subjects knew the location of the right answer, and hence before any preparation of the appropriate physical movement could have started.

That's quite a long way from mind-reading in the fullest sense. Only one distinction was tested - between a brain about to add and a brain about to subtract. It may well be that these are particular examples of a vast set of different patterns for all sorts of mental activity: but in strict logic it could be that this technique only allows us to distinguish between broadly augmentative or positive operations and diminutive or negative ones: it would be interesting to see whether multiplication and division, for example, are as readily distinguishable from addition and multiplication as the latter two are from each other. So far as I can tell, there is no evidence yet of any particular commonality, other than being in the same brain region, between the addition pattern in one individual and the addition pattern in another, or between the addition pattern in one subject today and the same thing in the same person next week.

But the results are a clear step forwards, and the fact that they open up many further questions is really a virtue, since a number of the further issues seem to be readily susceptible to further research. One of the more likely practical developments of the method, I imagine is the development of a superior lie detector (though it would need a higher success rate than 70%).

One of the early experiments which would be needed to firm up this possibility would presumably be to test whether the subjects can fake their own results. They would have to be asked to hold in their minds the intention of adding, say, while covertly knowing that they would in fact subtract. Would the result be the addition pattern, the subtraction pattern, or something different altogether?¶Considering that point takes us into some of the more confusing issues raised by the research. When the subjects were holding an 'intention to subtract' in their minds, what were they actually doing? There are a number of possibilities. They could have started up their mental subtraction engine, and be keeping it idling until the numbers arrived. They could have put up some kind of unconscious internal flag which merely indicated the intention, to be consulted when the task was performed. They could have been holding on to a conscious mental image of the relevant algebraic symbol, or the word 'subtract'. They could have been in a neutral mental state, relying on memory to retrieve the decision made a moment before.

These considerations may lead us to note the somewhat artificial nature of what the subjects were being asked to do, and consequently question whether the patterns identified in their brains really represent an intention, an awareness of an intention, a willed maintenance of an intention, the declaration of an intention, an awareness of a declaration of an intention, or whatever other higher-order possibilities we might come up with. Not that this makes the research any less interesting or significant, of course; rather the reverse.

The language used in the paper suggests an optimistic view, since it speaks of intentions being 'encoded' in the brain. Probably no particular commitment was intended by this and the word was being used in the sort of loose sense in which your shadow can be said to encode information about your height and shape. If there really were a decipherable mental code which was common to all brains, then we really could look forward to mind-reading in the fullest sense, but given the variation between brains, it may be that our individual ways of holding intentions have nothing much in common. Indeed, it may turn out that even on an individual basis our mental life is not organised in an encoded way. Computer code has to make sense on two levels: it has to be comprehensible to the programmer and deliver cogent outputs. The human brain is under no obligation to make sense except in output terms, and since rendering one's code understandable usually involves a small overhead, it's probable that evolution hasn't done so. If that pessimistic view is right, we're never going to get more than one-to-one matchings between neural activity and particular thoughts.

But that is the pessimistic view. It's a bit hard to believe that orderly thought could emerge from something that didn't have at least an implicit orderliness. Perhaps, to pursue the analogy, we can't expect the brain's code to include helpful comments; but surely efficiency will have led to some modularisation and categorisation of the kind which may give a foothold to interpretation?

If a clear answer about that emerges from further research, as it may, that certainly will be worth a big media write-up.'

Too Thin? Too Rich?

2 March 2007

Just before you'd read this sentence, were you consciously aware of your left foot? Eric Schwitzgebel set out to resolve the question in the latest edition of the JCS.

In normal circumstances, we are bombarded by impressions from all directions. Our senses are constantly telling us about the sights, sounds and smells around us, and also about where our feet are, how they feel in their shoes, how hungry we currently feel, whether our sore calf muscle is any better at the moment; and what spurious reasoning some piece of text on the internet is trying to spin for us. But most of the time, most of this information is ignored. In some sense, it's always there, but only the small subset of things which are currently receiving our attention are, as it were, brightly lit.

There's little doubt about this basic scenario. Notoriously, when we drive along a familiar route, the details drop into the background of our mind and we start to think about something else. When we arrive at our destination, we may not remember anything much about the journey: but clearly we could see the road and hear the engine at all relevant times, or we probably shouldn't have been able to finish the journey. On the other hand, suppose as we were driving along, the sound of a baby crying had unexpectedly drifted over from the back seat: would we have failed to notice that feature of the background?

So we have two (at least two) levels of awareness going on. Schwitzgebel poses the question: which do we regard as conscious? On the "thin" view, we're only really conscious of the things we're explicitly thinking about. No doubt the other stuff is in some sense available to consciousness, and no doubt bits of it can pop up into consciousness when they trigger some unconscious alarm; but it's not actually in our consciousness. How else, the thinnists might ask, are we going to make the distinction between the two different levels? The rich view is that everything should be included: I may not be thinking about my foot at all times, but to suggest that I only know where it is subconsciously, or unconsciously, seems ridiculous.

Schwitzgebel does not think either side has particularly strong arguments. Both are inclined to provide examples, or assert their case, and expect the conclusion to seem obvious. Searle has argued that we couldn't switch attention unless we were conscious of the thing we were switching our attention to: Mack and Rock have done experiments to prove that while paying close attention to one things we may fail to notice other things: but neither of these lines of discussion really seems to provide what you call a knock-down case.

Accordingly, with many reservations, Schwitzgebel set up an experiment of his own. The subjects wore a beeper which went off at a random period up to an hour after being set: they then recorded what they were conscious of immediately beforehand (it's important, of course, to keep the delay minimal, otherwise the issue gets entangled with problems of memory). The subjects were divided into groups and asked to focus on tactile, visual and total sensory experience.

The results supported neither the thin nor the rich position. Perhaps the most interesting finding is the degree of surprise evoked in the subjects. In a departure from normal experimental

method, Schwitzgebel used as subjects philosophy postgrads who could reasonably be expected to have some established prejudices in the field: he also spent time explaining the experiment and talking over the issues, and recorded whether each subject was a thinnist or richist at the start. Although this involved some risk of skewing the results, it allowed the discovery that thinnists actually often found themselves having rich experience, and vice versa.

Where does that leave us? It seems almost as though the dilemma is merely reinforced: the research points towards some compromise, but it's hard to see where we can find room for one. The results did seem to reinforce the existing general agreement that there really are two distinct levels or aspects of consciousness at work. Wouldn't one solution, then, be to give both neutral labels (con-1 and con-2?) and leave it at that? That may be what one of Schwitzgebel's subjects, who apparently dismissed the whole thing as 'linguistic' had in mind. But it's not a very comfortable position to dismiss the concept of consciousness in favour of two hazy new ones. Schwitzgebel, rightly, I think, considers that the difference between thinnism and richism is real and significant.

My best guess for a neatish answer is that we're simply dealing with pure first order consciousness and the same thing combined with second order consciousness. In other words, the dim constant awareness of everything being reported by our senses, really is conscious, but it's a region of consciousness we're not conscious of being conscious of. By contrast, we're not only conscious of the things at the forefront of our minds, we're also aware of being conscious of them. (It might well be that second-order consciousness is what animals largely or wholly lack - I wonder if thinnists also tend to be sceptics about animal consciousness?)¶Alas, that's not really a compromise: it seems to make me a kind of richist.

What Is This Thing Called Love?

13 March 2007

Edge has excerpted the first chapter of Marvin Minsky's book on emotions, The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. I clearly need to read the whole book, but I found the excerpt characteristically thought-provoking. As you might have expected, the general approach builds on the Society of Mind: emotions, it seems, allow us to activate the right set of resources (their nature deliberately kept vague) from the diffuse cloud available to us. So emotions really serve to enhance our performance: in young or unsophisticated organisms they may be a bit rough and ready, but in adult human beings they are under a degree of rational control and are subject to a higher level of continuity and moderation.

At first sight, explaining the emotions by identifying the useful jobs they do for us seems a promising line of investigation. Since we are the products of evolution, it seems a good hypothesis to suppose that the emotional states we have developed must have some positive survival value. The evidence, moreover, seems to support the idea to some degree: anger, for example, corresponds with physiological states which help get us ready for fighting. Love between mates presumably helps to establish a secure basis for the production and care of offspring. The survival value of fear is obvious.

However, on closer examination things are not so clear as they might be. What could the survival value of grief be? It seems to be entirely negative. Its physiological manifestations range from the damaging (loss of concentration and determination) to the surreal (excessive water flowing from the tear-ducts down the face). Darwin himself apparently found the crying of tears 'a puzzler' with no practical advantages either to modern humans or to any imaginable ancestor, or any intelligible relationship to any productive function - what has rinsing out your eyes got to do with the death of a mate or child? If it comes to that, even anger is not an unalloyed benefit: a man in the grip of rage is not necessarily in the best state to win an argument, and it's surely even debatable whether he's more likely to win a fight than someone who remains rational and judicious enough to employ sensible tactics.

One of the thoughts the extract provoked in me, albeit at a tangent to what Minsky is saying, concerned another possible problem with evolutionary arguments. The physiological story about an increased pulse rate and the rest of it is one thing, but does all that have to be accompanied by feeling angry? Can't we perhaps imagine going through all the right physiological changes to equip us for fighting, fleeing, or whatever other activity seems salient, without having any particular feelings about it?This sounds like qualia. If emotions are feelings which are detachable from physical events, are they, in themselves, qualia? I'm not quite sure what the orthodox view of this is: I've read discussions which take emotions to be qualia, or accompanied by them, but the canonical examples of qualia - seeing red and the rest of it - are purely sensory. The comparison, at any rate, is interesting. In the case of sensory qualia there are three elements involved: an object in the external physical world, the mechanical process of registration by the senses, and the ineffable experience of the thing in our minds, where the actual redness or sounding or smelliness occurs. In the case of the emotions, there isn't really any external counterpart (although you may be angry or in love with someone, you perceive the

anger or love as your own, not as one of the other person's qualities): the only objective correlate of an emotional quale is our own physiological state.

Are emotional zombies really possible? It's widely though not universally believed that we could behave exactly the way we do - perhaps be completely indistinguishable from our normal selves - and yet lack sensory qualia altogether. An emotional zombie, along similar lines, would have to be a person whose breathing and pulse quickened, whose face blushed and voice turned hoarse, and who was objectively aware of these physiological manifestations, but actually felt no emotion whatever. I think this is still conceivable, but it seems a little stranger and harder to accept than the sensory case. I think in the case of an emotional zombie I should feel inclined to hypothesise a kind of split personality, supposing that the emotions were indeed being felt somewhere orin some sense, but that there was also a kind of emotionless passenger lodged in the same brain. I don't feel similarly tempted, with sensory zombies, to suppose that real subjective redness must be going on in a separate zone of consciousness somewhere in the mind.

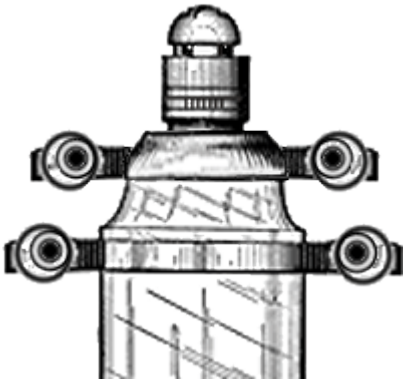
The same difference in plausibility is visible from a different angle. In the case of sensory qualia, we can worry about whether our colour vision might one day be switched, so that what previously looked blue now looks yellow, and vice versa. I suppose we can entertain the idea that Smith, when he goes red in the face, shouts and bangs the table, is feeling what we would call love; but it seems more difficult to think that our own emotions could somehow be switched in a similar way. The phenomenal experience of emotions just seems to have stronger ties to the relevant behaviour than phenomenal experience of colours, say, has to the relevant sensory operations.

It might be that this has something to do with the absence of an external correlate for emotions, which leaves us feeling more certainty about them. We know our senses can mislead us about the external world, so we tend to distrust them slightly: in the case of emotions, there's nothing external to be wrong about, and we therefore don't see how we could really be wrong about our own emotions. Perhaps, not entirely logically, this accounts for a lesser willingness to believe in emotional zombies.

Or, just possibly, emotional qualia really are tied in some deeper way to volition. This is not so much a hypothesis as a gap where a hypothesis might be, since I should need a plausible account of volition before the idea could really take shape. But one thing in favour of this line of investigation is that it holds out some hope of explaining what the good of phenomenal experience really is, something lacking from most accounts. If we could come up with a good answer to that, our evolutionary arguments might gain real traction at last.

What, No Kill-Bots?

31 March 2007



You may have read a month or two ago about the rather scary robotic sentries which have been created: it seems they identify anything that moves and shoot it. Although it has a number of interesting implications, that technology does not seem an especially exciting piece of cybernetics. The BICA project (Biologically Inspired Cognitive Architectures) set up by DARPA (the people who, to all intents and purposes, brought us the Internet) is a very different kettle of fish.

The aim set out for the project was to achieve artificial cognition with a human level of flexibility and power, by bringing together the best of computational approaches and recent progress in neurobiology. Ultimately, of course, the application would be military, with robots and intelligent machines supporting or superseding human soldiers. In the first phase, a number of different sub-projects explored a range of angles. To my eyes there is a good deal of interesting stuff in the reports from this stage: if I had been sponsoring the project, I should have been afraid that each team would want to go on riding its own favourite hobby-horse to the exclusion of the project's declared aims, but that does not seem to have happened.

In the second phase, the teams were to proceed to implementation, and the resulting machines were to compete against each other in a Cognitive Decathlon. Unfortunately, it seems there will be no second phase. No-one appears to know exactly why, but the project will go no further.

It could well be that the cancellation is the result of budget shifts within DARPA that have little or nothing to do with the project's perceived worth. Another possibility is that the sponsors became uneasy with the basic idea of granting lethal hardware a mind of its own: the aim was to achieve the kind of cognition that allows the machine to cope with unexpected deviations from plan, and make sensible new decisions on the fly; but that necessarily involves the ability to go out of control spontaneously. It could also be that someone realised how difficult moving from design to implementation was going to be. It has always been easy to run up a good-looking high-level architecture for cognition, with the real problems having a tendency to get tidied up into a series of black boxes. This might have been one project where it was a mistake to start with the overall design. The plasticity of the human brain, and the existence of other brain layouts in creatures such as squid, suggest that the overall layout may not matter all that much, or at least that a range of different designs would all be perfectly viable if you could get the underlying mechanisms right.

There is another basic methodological issue here, though. When you start a project, you need to know what you're trying to build and what it's supposed to do: but no-one can really give a clear answer to those questions so far as human cognition is concerned. The BICA project was likened by some to the Apollo moon landings: but although the moon trips were hugely challenging, it was always clear what needed to be delivered, and in broad terms, how.

But what is human cognition actually for? We can say fairly clearly what some of the sub-systems do: analyse input from the eyes, for example, or ensure that the sentences we utter hang together properly. But high-level cognition itself?

From an evolutionary perspective, cognition clearly helps us survive: but that could equally be said of almost every organ and function of a human being, so it doesn't help us define the distinctive function of thought. Following the line adopted by DARPA we could plausibly say that cognition frees us from the grasp of our instincts, helping us to deal much more effectively with novel situations, and exploit opportunities which would otherwise be neglected. But that doesn't really pin it down, either: the fact that thoughtful behaviour is different from instinctive, pre-programmed behaviour doesn't distinguish it from random behaviour or inertia, and pointing out that it's often more successful behaviour just seems to beg the question.

It seems to be important that human-level cognition allows us to address situations which have not in fact occurred; we can identify the dangerous consequences of a possible course of action without trying it out, and enable 'our hypotheses to die in our stead'. Perhaps we could describe cognition as another sense, the sense of the possible: our eyes allow us to consider what is around us, but our thoughts allow us to consider what would or might be. It's surely more than that, though, since our imagination allows us to consider the impossible and the fantastic just as readily as the possible. As a definition, moreover, it's still not much use to a designer, not least because the very concept of possibility is highly problematic.

Perhaps after all we were getting closer to the truth with the purely negative point that thoughtful behaviour is not instinctive. When evolution endowed us with high-level cognition, she took an unprecedented gamble; that cutting us loose to some degree from self-interested behaviour would in the end and overall lead to better self-interested behaviour. The gamble, so far, appears to have paid off; but just as the kill-bots could choose alternative victims, or perhaps become pacifists, human beings can (and do) kill themselves or choose not to reproduce. Perhaps the distinctive quality of cognition is it free, gratuitous character: its point is that it is pointless. That doesn't seem to be much help to an engineer either.

Anyway, I think I can wait a bit longer for the kill-bots; but it seems a shame that the project didn't go on a bit further, and perhaps illuminate these issues.

Down With Physics

14 April 2007

Robert Lanza, well-known in the area of cloning and stem cells, has fired off a broadside in another direction. Never mind the physicists, he says, with their long-awaited Theories of Everything and their bizarre multi-dimensional intangible substrates: in fact the fundamental science is biology. Space and time are not even real except inasmuch as we perceive them; and perception is a product of consciousness, a biological mystery the physicists can never hope to penetrate.

It's easy to sympathise with Lanza's irritation over the triumphalism indulged in by some physicists - while physics itself seems in some ways to be getting ever more deeply into difficulty. In terms of clear progress and perhaps even methodological purity, biology seems to have a far better story to tell in recent years. But can biology really be more fundamental than physics?

Lanza's key theme is that reality depends on perception; perception on consciousness; and consciousness on biology. With the first step, we're back with Bishop Berkeley, whose controversial view that 'to be is to be perceived' Lanza almost seems to take for granted. A substantial, centuries-long debate has already taken place over this, which I cannot hope to do justice to here; but to take just one contrary argument: if all reality depends on perception, there's an unsolved problem about how things get started at all. There I am, for the sake of argument, hanging in the metaphysical void. Nothing exists until I perceive it; but how do I start perceiving something which doesn't exist? Even my own thoughts must be there before I can become aware of them; yet they can't exist until I have perceived them. So I can't even think? Lanza acknowledges that some have seen his philosophy as leading inevitably to solipsism: but it seems it might lead to utter nullity. Lanza says that the illusions suffered by schizophrenic patients are as real to them as the ordinary world is to us: but the possibility of error is not sufficient to demonstrate the impossibility of truth (though Lanza is not the first person to have given up too easily on objective reality). In places Lanza actually seems rather equivocal about his Berkeleyanism: he offers the analogy of a CD player: until it works on the relevant tracks, the music doesn't exist: and in the same way, there's no reality until our minds have operated on... what? It ought to be the underlying realities of physics, but they are what Lanza seems to want to deny.

Though no doubt it is true that consciousness is biological, that cannot altogether be taken for granted either. Among others Lanza cites Descartes, Kant, and Leibniz as well as Berkeley in support of the primacy of consciousness: but none of them would have accepted that it was a matter of biology. When Hume daringly had one of his characters declare that the processes of consciousness in the brain were not fundamentally different from the processes of decay in a cauliflower, he took care to distance himself from a view he knew would be regarded as an insult to the human spirit, far beyond the pale of civilised discourse. Moreover, Lanza's own views, curiously enough, place a barrier between him and the biology he wishes to celebrate. Our knowledge of biology, after all, comes to us in much the same kind of way as our knowledge of physics; through our senses - our perceptions. If time and space, and other concepts of physics, are really illusions, then surely so are cells and organisms and brains. Lanza says experience is something generated 'inside your head', but what head? The only knowledge we have of heads comes to us through, guess what, experience. The truth here seems to be that

biology simply cannot take the role Lanza wants to assign to it, and in seeking to 'get below' physics, he ends up resorting to metaphysics, the only subject (with the possible exception of maths) which really does operate at an even more fundamental level.

Lanza puts forward a couple of other arguments to support his case. He appeals (curiously enough) to physics itself in the shape of quantum theory, which he suggests has eliminated the idea of a reality independent of perception. Like Berkeleyanism, the correct interpretation of quantum physics is a large subject, but my impression is that it would be at the radical end of the spectrum to suppose that it did away with the idea of an objective, independent reality altogether.

He also mentions the argument that many features of the universe seem to have been set up with great precision to allow the eventual possibility of life. If gravity or the strong nuclear force were slightly different from what they are, the world would never have been a habitable place. I'm not exactly certain about how Lanza means this to fit withhis overall view, but it seems he must be suggesting that the world actually began, in some non-chronological sense, with human perception, which then extrapolated backwards the necessary conditions for its arising in a presumed physical world. If so, I'd like more explanation about how that would work and why it would necessarily give rise to these exquisitely precise physical constants: but the underlying anthropic argument seems weak to me.

First of all, the reasoning smacks of Warty Bliggens, the toad:

' he explained that when the cosmos was created that toadstool was especially planned for his personal shelter from sun and rain thought out and prepared for him'

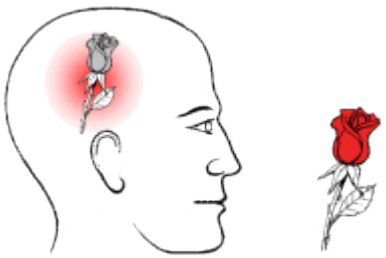
It's not surprising that the set-up of the Universe favours the existence of human beings because if it didn't, we shouldn't be here to worry about it (but perhaps something else would).

It may be, in fact I suspect it must be, that there are, as yet unknown to us, deep metaphysical reasons why the cosmos is the way it is and could not have been otherwise: in which case there's no real scope for surprise about the way it turned out. But if the basic laws and constants are in some way arbitrary, as Lanza's argument supposes, we can't really claim to know what the range of possible universes, or the range of possible conscious entities, really is. It may be that tinkering slightly with a few of the current constants produces a world in which human beings cannot occur; but why stop there? If we vary the basic rules more fundamentally, it might well be that there are countless possible universes utterly unlike ours, containing innumerable multitudes of unimaginable thinking beings. In that case, it is again unsurprising that we should chance to occur in a possible world which happens to suit us.

So although it's interesting to entertain the idea, I don't think biology, for all its merits, can be enthroned as the most fundamental of sciences.

Colour Blind Colours

22 April 2007



A team of researchers (Milán, Hochel, González, Tornay, McKenney, Díaz Caviedes, Mata Martín, Rodríguez Artacho, Domínguez García and Vila) have a paper in the JCS setting out their experiments with an unusually interesting subject. This person, referred to as 'R', is colour blind for certain colours: but he also experiences a spectrum of synaesthetic colours in which he can make the full range of clear distinctions. The team believes its research shows that qualia can be accessible and useful to science.

Synaesthesia, the occurrence of sensations (often sensations of colour) in association with unrelated stimuli, is an interesting but treacherous subject. Many people feel that the days of the week, or the numbers from one to ten, seem to have their own appropriate colours, and sceptics may feel that the whole phenomenon of synaesthesia is largely a case of people allowing their poetic feelings to run away with them. But there is a good deal of evidence to show that vivid synaesthesia is a real phenomenon. One of the earliest and most striking of reported cases is the classic one recounted by Luria in The Mind of a Mnemonist. His subject (referred to as 'S') formed exceptionally strong associations, which had the benefit of giving him a 'photographic' memory but was also a considerable handicap as unwanted associations intruded into his life, making it impossible, for example, to buy an ice cream when something in the vendor's voice called up a vivid impression of red hot coals.

R's synaesthesia is not on this troublesome scale: he experiences colour sensations in connection with pictures and people, but he experiences the colour as internal, rather than seeing it 'out there' as some synaesthetes do. For him, there is a kind of code, with each synaesthetic colour having certain emotional connotations: in the case of people, for example, red means attractive, green means ill or dirty, purple means upbeat, yellow means envious or aggressive, and brown means old or uninteresting (the reader may be able to guess at R's own approximate age). A portion of the team's research was devoted to exploring these emotional connotations, and they claim, unconvincingly I think, that their results show the possibility of an inverted spectrum of emotions (I think they merely show that R's emotional reactions are in some respects atypical: the sky for him evokes red, which makes it exciting, whereas it is more commonly seen as restful: but that doesn't amount to an inverted emotional spectrum).

The exploration of synaesthesia in colour blind subjects is not new - Ramachandran and Hubbard have produced many interesting findings, including a subject who referred to colours he could experience synaesthetically, but never with his eyes, as 'Martian' colours. This naturally raises the question of whether the Martian colours were ones which people with normal vision are used to, or something genuinely stranger. How would we know?

The exploration of qualia recounted in the JCS piece rests on an ingenious use of the Stroop effect. This effect occurs where, for example, the subject is shown a series of colour words, and asked to say, not what colour they name, but what the colour of the font is. If you've tried this, you'll know that it is more difficult to do quickly and accurately than you might suppose: much harder, in any case, than when font and word indicate the same colour. Now if you were completely colour blind, you would of course be immune to Stroop interference. The team

devised experiments which showed that R did indeed fail to show Stroop effects in certain cases where other subjects suffered them, very plausibly because of his limited colour vision. However, when he was shown pictures which evoked dissonant synaesthetic colours, Stroop interference did occur.

This seems to show that R's inner colour experience - his qualia? - cover the normal range, distinct from the range of colours he can actually distinguish with his eyes. Moreover, in a separate series of experiments, the team got him to match his synaesthetic colours with real ones, confirming that in his case at least they were not Martian in character. Does all this show that qualia are 'a useful scientific concept'?

Interestingly, the paper quotes (and slightly misdescribes) one of the 'intuition pumps' used by Daniel Dennett. Two coffee tasters (Chase and Sanborn) have ceased to enjoy Maxwell House, but seemingly for different reasons. To one, it tastes the same as ever, but he no longer likes that taste: the other still likes that taste, but the coffee, although it has not changed chemically or in any other objective respect, no longer tastes that way to him. Dennett's case is that this distinction, between the qualia and the taster's reaction, is not ultimately sustainable. One of the taster's wives ultimately straightens him out on the point: so he's still having the same taste experience, but now doesn't like it? But doesn't the fact that he doesn't like it mean, in itself, that the experience is different? There's just no point in talking about qualia as distinct from our reactive dispositions.

What would Dennett say about R? I think he would say the research is an interesting exploration of those reactive dispositions, but that nothing here requires us to talk of qualia at all. R's brain reacts to certain stimuli in ways which other people's brains react to certain inputs from their eyes, though R himself lacks those inputs. But just because synaesthetic colour doesn't come from the eyes, we mustn't conclude that it has the inner, subjective quality of qualia. Indeed, by ingeniously naturalising synaesthetic colour, and showing it to be accessible to science, the team has arguably shown that it can't be identified with qualia.

The team concludes that their research shows the coffee taster thought experiment actually makes no sense: the quale and the reaction 'seem to be rigidly connected and cannot change independently'. Dennett might not be unhappy with that conclusion: why not take one more step, he might say, and draw the conclusion that talk of qualia is really only talk of reactions...?

Loopy

1 May 2007

With I am a Strange Loop, Douglas Hofstadter returns - loops back? to some of the concerns he addressed in his hugely successful book Gödel, Escher, Bach, a book which engaged and inspired readers around the world. Despite the popularity of the earlier book, Hofstadter feels the essential message was sometimes lost; the new book started out as a distilled restatement of that message, shorn of playful digressions, dialogues, and other distractions. It didn't quite turn out that way, so the new book is much more than a synopsis of the old. However, the focus remains on the mystery of the self.

Is it a mystery? In a way I wondered why Hofstadter thought there was any problem about it. Talking about myself is just a handy way of picking out a particular human animal, isn't it? In fact it doesn't have to be human. We might say of a dog that "He wants to get in the basket himself"; for that matter we might say of a teacup "It just fell off the shelf by itself". Why should talk of I, myself, me, evoke any more philosophical mystery than talk of she, herself, her?

I can think of three bona fide mysteries attached to the conscious self: agency (or free will if you like); phenomenal experience (or qualia); and intentionality (or meaning). Hofstadter is unimpressed by any of these. Qualia and free will get a quick debunking in a late chapter: Hofstadter shares his old friend Dennett's straight scepticism about qualia, and as for free will, what would it even mean? Intentionality gets a more respectful, but more glancing treatment Hofstadter proposes the analogy of the careenium, which is to be imagined as kind of vast, frictionless billiard table on which hordes of tiny magnetic balls, know as simms, roll around. Occasionally the edge of the careenium may be hit by objects in the exterior world, imparting new impulses to the simms and possibly causing them to agglomerate into big rolling masses, or simmballs (the attentive reader will notice that a strong focus on the core message has by no means excluded a fondness for artful puns and wordplay). Evidently impulses from the outside world spark off neural symbols, whose interaction gives rise to our thoughts.

It would not be fair to criticise the sketchiness of this account, because it isn't what Hofstadter is on about: it isn't even the main point of the careenium metaphor. He acknowledges, too, that some will think talk of symbols in the brain leads to the assumption of an homunculus to read them, and to other difficulties, and he briefly indicates a response. The point really is that for Hofstadter all this is largely beside the point.

So what is the real point, and why does Hofstadter find the self worthy of attention? For him it is all about greatness of soul; friendship and the sharing, the sympathetic entering into, of other consciousnesses. This is where the loops come in.

Hofstadter quotes several examples of self-referential loops, re-creating, for example, experiments with video feedback which he first carried out many years ago. More substantially, he gives an account of how Kurt Gödel managed to get self-reference back into logical systems like the one set out in Russell and Whitehead's Principia Mathematica. PM, as Hofstadter calls it, was a symbolic language like a logical algebra, which the authors used to show how

arithmetic could be built up out of a few simple logical operations. One of its distinguishing features was that it included special rules which were meant to exclude self-reference, because of the paradoxes which readily arise from sentences or formulae that talk about themselves (as in 'This sentence is false.'). By arithmetising the notation and applying some clever manoevres, Gödel was able to reintroduce self-reference into the world of PM and similar systems (a disaster so far as the authors' aspirations were concerned) and incidentally went on to prove the existence in any such system of true but unprovable assertions.

This is one of the most interesting sections of the book, though not the easiest to read. It's certainly enlivened, however, by the grudge Hofstadter seems to have against Russell: I found myself wondering whether Grammaw Hofstadter could have been the recipient of unwelcome attentions from the aristocratic logician at some stage. I've read many accounts which suggest that Bertrand Russell might have been, say, a touch self-centred: but it's a novelty to read the suggestion that he might have been a bit dumb in certain respects. In this account, Gödel is the young Turk who explodes the enlivening force of self-reference within Russell's gloomy castle: no mention here of how Russell himself had previously done something similar to the structure erected by Frege, so that Russell also has a claim to be considered a champion of paradox.

People are clearly examples of self-referential systems, but is self-reference an essential part of consciousness or merely a natural concomitant? It is plausible that self-reference might help explain why our thoughts, for example, seem to pop out of nowhere: when we try to locate the source, we get trapped into looking down a regress like the infinite corridor which appears in the video feedback. Hofstadter's loops, moreover, are no ordinary loops - they are strange. A strange loop violates some hierarchical order, perhaps with containers being contained, or steps from a higher level leading up to lower ones. It is our symbolic ability which gives us the ability to create strange loops: to symbolise our own symbolic system, and so on. However, we are unable to perceive the lower-level neural operations which constitute our thoughts, and we therefore suffer the impression of mysteriously efficacious, self-generating thoughts and decisions, which we cannot reconcile with the sober scientific account of the physical events which really carry the force of causality.

Hofstadter spends some time on this issue of different levels of interpretation, seeking to persuade us that the existence of complete physical or neural stories does not stop accounts of causality in terms of higher-level entities (thoughts, intentions, and so on) from being salient and reasonable. He argues, intriguingly, that there is an analogy here with what Gödel did: by reinterpreting PM at a higher level, he was able to draw conclusions and in a sense set limits to what could be done within PM. In the same way we can legitimately say that our intentions were pushing the neurons around, not just that the neurons were pushing the intentions. This idea of downward causality across levels of interpretation seems metaphysically interesting, though I think it is quite redundant: the relation between entities at different levels of explanation is identity, not causation (though I suppose you can see identity as the basic case of causation if you wish).

Now things get more difficult. Hofstadter argues that bits of our own loops get echoed in those of other people, and that therefore we exist in them as well as in our own brains. If two video cameras and screens are looping and enter each other's view, then the loop of camera A will be running, in miniature, in that of camera B, and vice versa. Our engagement with other people thus helps to create and sustain us, and in fact it means that death itself is not as abrupt as it would otherwise be, the afterglow of our own strange loops continuing to spin in the minds of others and thereby sustaining our continued existence to some limited extent. This theorising is

attached to an account of Hofstadter's reaction to the sudden, and tragically early, death of his wife.

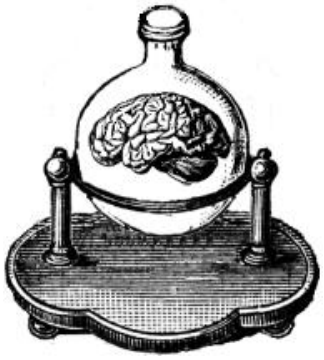
Hofstadter does not believe that his views about death have been fundamentally influenced by this awful experience: he says that his conclusions were largely drawn before his wife's sudden illness. All the same, the context makes it more difficult to say that, in fact, the reasoning here is wholly unconvincing. It's just not true that the image of a feedback loop in another feedback loop remains a viable feedback loop in itself.

Hofstadter offers some arguments designed to shake our belief in the simple equation of body with person (he is happy to use the word 'soul', though without its usual connotations), including the one - rather tired in my view - of the magic scanners which can either relocate or duplicate us perfectly. Where a person is duplicated like this, he says, it makes no sense to insist that one of the two bodies must be the real person. "…to believe in such an indivisible, indissoluble "I" is to believe in nonphysical dualism" . Not necessarily: it might instead be to assert the most brutal kind of physical monism: we are physical objects, and our existence depends utterly on our physical continuity.

I'm afraid I therefore end up personally having a good deal of sympathy with many of the incidental things Hofstadter says while remaining unconvinced by the points he regards as most important. Whether or not you accept his case, though, Hofstadter has a remarkable gift for engaging and lively exposition, more of which will always be welcome.

I Ain't Got No Body

13 May 2007



You probably read about the recent experiment which apparently simulated enough neurons for half a mouse brain running for the equivalent of one second of mouse time - though that required ten seconds of real time on a massively powerful supercomputer. Information is scarce: we don't know how detailed the simulation was or how well the simulated neurons modelled the behaviour of real mouse neurons (it sounds as though the neurons operated in a void, rather than in a simulated version of the complex biological environment which real neurons inhabit, and which plays a significant part in determining their behaviour). No attempt was made to model any of the structure of a mouse brain, though it is said that some signs of patterns of firing were detected, suggesting that some nascent self-organisation might have been under way - on an optimistic interpretation.

I don't know how this research relates to the Blue Brain project, which has just about reached what the researchers consider a reasonable neuron-by-neuron electronic simulation of one of the cortical columns in the brain of a juvenile rat. The people there emphasise that the project aims to explore the operation of neuronal structures, not to produce a working brain, still less an artifical consciousness, in a computer.

It is not altogether clear, of course, how well an isolated brain - even a whole one - would work (another unanswered question about the simulation is whether it was fed with simulated inputs or left to its own devices). A brain in a jar, dissected out of its body but still functioning, is a favourite thought-experiment of some philosophers; but others denounce the idea that consciousness is to be attributed to the brain, rather than the whole person, as the 'mereological fallacy'. A new angle on this point of view has been provided by Rolf Pfeifer and Josh Bongard in How the body shapes the way we think. Their title is actually slightly misleading, since they are not concerned primarily with human beings, but seek instead to provide a set of design principles for better robots, drawing on the history of robotics and on their own theoretical insights. They're not out to solve the mystery of consciousness, either, but their ideas are of some interest in that connection.

In essence, Pfeifer and Bongard suggest that many early robots relied too much on computation and not enough on bodies and limbs with the right kind of properties. They make several related points. One is the simple observation that sometimes putting springs in a robot's legs works better than attempting to calculate the exact trajectory its feet need to follow to achieve perfect locomotion. In fact, a great deal can be accomplished merely by designing the right kind of body for your robot. Pfeifer and Bongard cite the 'passive dynamic walkers' which achieve a human-style bipedal gait without any sensors or computation at all (they even imply, bizarrely, that the three patron saints of Zurich might have worked on a similar principle: apparently legend relates that once their heads were cut off, the holy three walked off to the site of the Grossmünster church). Similarly, the canine robot Puppy produces a dog-like walk from a very simple input motion, so long as its feet are allowed to slip a bit. Human babies are constructed in such a way that even random neural activity is relatively likely to make their hands sweep the area in front of them, grasp an object, and bring it up towards eyes and mouth: so that this exploratory behaviour is inherently likely to arise in babies even if it is not neurally pre-programmed.

Another point is that interesting (and intelligent) behaviour emerges in response to the environment. Simple block-pushing robots, if designed a certain way, will automatically push the blocks in their environment together into groups. This behaviour, which is really just a function of the placement of sensors and the very simple program operating the robots, looks like something achieved by relatively complex computation, but really just emerges in the right environment.

Thins are looking bleak for the brain in the jar, but there is a much bolder hypothesis to come. Pfeifer and Bongard note that a robot such as Puppy may start moving in an irregular, scrambling way, but gradually falls into a regular trot: in fact, for different speeds there may be several different stable patterns: a walk, a run, a trot, a gallop. These represent attractor states of the system, and it has been shown that neural networks are capable of recognising these states. Pfeifer and Bongard suggest that the recognition of attractor states like this represents the earliest emergence of symbolic structures; that cognitive symbol manipulation arises out of recognising what your own body is doing. Here, they suggest, lies the solution to the symbol grounding problem; of intentionality itself.

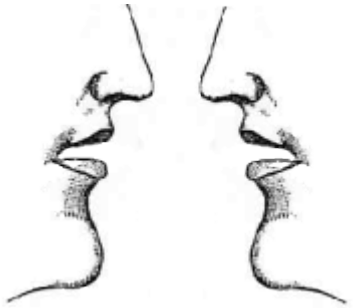
If that's true, then a brain in a jar would have serious problems: moreover, simulating a brain in isolation is very likely to be a complete waste of time because without a suitable body and an environment to move around in, symbolic cognition will just never get going.

Is it true? The authors actually offer relatively little in the way of positive evidence or argument: they tell a plausible story about how cognition might arise, but don't provide any strong reasons to think that it is the only possible story. I'm not altogether sure that their story goes as far as they think, either. Is the ability to respond to one's body falling into certain attractor states a sign of symbol manipulation, any more than being able to respond to a flash of light or a blow on the knee? I suspect that symbols require a higher levl of abstraction, and the real crux of the story is in how that is achieved - something which looks likely, prima facie, to be internal to the brain.

But I think if I were running a large-scale brain simulation project, these ideas might cause me some concern.

Haecceity and metempsychosis'

3 June 2007



I keep an open mind on the subject of qualia, the ineffable redness of red and the equally ineffable smelliness of smell, and I always read new arguments on the subject with interest. But if I'm honest, in my heart of hearts I believe that the problem, as normally formulated, embodies a misunderstanding. That redness you see isn't a mysterious quale; it's just some real, extra-theoretical redness. When Mary sees something red for the first time in her life, she doesn't gain new information or knowledge, she just has an experience she has never had before. We should not be puzzled over the fact that colour theories do not themselves contain real colours, and so cannot convey them into minds any more than knowledge about an experience conveys the experience itself.

I think we are right to be puzzled, nevertheless. There is a profound mystery here - it just isn't the one we usually discuss. The redness of an apple is just real redness: but what on earth is that? What is real anything? When we try to distinguish reality from dreams, we often rely on epistemic arguments about the consistency and coherence of the real world, but those properties are not of the essence: there's nothing to say a dream or delusion couldn't by chance be consistent and sustained. The realness of reality must lie somewhere else, in some inscrutable region of meta-ontology.

Part of the problem, at least, is what we might call haecceity (with no particular commitment to Duns Scotus or others who have used the word): thisness: the property of being a particular thing and not something else. One of the deepest metaphysical questions we can ask is why everything is like this, instead of like something else: indeed, why is everything like anything, instead of nothing (which after all would require much less explanation)? Abstractions may sometimes be hard to get your head around, but when you come down to it there's nothing so incomprehensible as the detailed particularity of ordinary reality.

We might look for some kind of cosmogonical explanation for why the world is the way it is. Now that we all know about chaos and have played with Conway's Game of Life, it seems relatively plausible that quite a simple starting point for the Universe could have given rise to the buzzing complexity of the real world. To be real, then, would be to feature in the unrolling of some cosmic algorithm. But this seems unsatisfactory in more than one respect. First, the ultimate explanation is now deferred to the laws or rules which sustain the algorithm, whether they are simply the laws of physics or some unknown principles of maths, logic, or metaphysics. Apart from the fact that we now have the new and daunting problem of explaining how and why these laws work, we're back with the old problem of putting the source of reality in a theory, the same sort of mistake we were making with qualia in the first place. Speaking for myself, I see theories and laws of nature as explaining and clarifying the innate tendencies of stuff: to say that the reality of the stuff arises out of the operation of the theory, or the laws, seems very like a category mistake.

Moreover, since it is a real thing, consciousness also has haecceity; but I find it impossible to believe that my own consciousness is merely a product of some rolling program: it may seem like a feeble recourse to intuition, but that just doesn't seem a satisfying answer. It doesn't

explain why I am so definitely here and not there, or so definitely this person and not that one. I am not a zombie, in the philosophical sense, and my existence has a more solid character than any Platonic process. And let's face it, it is all about me. I'm keen to understand the nature of reality in general, but like Descartes, I want my own reality to be clarified first and above all.

Perhaps metempsychosis is part of the answer. A post of Shankar's a while back, drawing on Hindu ideas, proposed that qualia reside in a universal background entity, which gets plugged into the disparate personalities and bodies of all of us. I think some tricky stuff needs to be put in place to deal with that kind of split, but the idea of one soul, as it were, doing the work for all of us, is interesting. For one thing it clears up the tricky housekeeping side of reincarnation. Since deaths and births are not neatly paired up, reincarnation requires at least a kind of celestial waiting-room, and possibly a kind of time-travelling on the part of the transmigrating souls. But if one soul can animate an indefinite number of people simultaneously, the problem doesn't arise. And why shouldn't things work like that? That magic core which sustains my selfhood doesn't seem to be dependent on the details of my body, my situation in history, or perhaps even my personality, when you get right down to it. So it doesn't seem to have many differentiating characteristics. If it's indistinguishable from other people's magic core, then perhaps a rather slapdash application of Leibniz's Law suggests it's really the same. Perhaps in some sense all human experiences are experienced by the same essential me, either sequentially or somehow simultaneously; perhaps, in fact, I am talking to myself here? So I can conclude that it would indeed be odd and puzzling if I were an island of haecceity - because in fact I am nothing less than universal subjectivity.

Convinced? No, nor me, I'm afraid. In the end all that we're left with, at best, is a persuasive denial of the very haecceity, so pressingly evident, which I wanted explained in the first place. I'm not convinced that my core of selfhood can be stripped of all its characteristics without losing its identity, and if it could that would identify me, not just with other people, but with everything. To some that might be a congenial conclusion, but to me again it looks as if the identity we wanted elucidated has run through our fingers. It seems more likely to me that my personality and the quirks of my physical history are constitutive of what I am even in the deepest sense: unfortunately of course, those quirky particularities are as hard to account for as my own thisness.

So it's back to the metaphysical drawing board, I fear. Perhaps the next set of ideas about qualia will, after all, hold the vital clue?

Animal Problems

10 June 2007

The latest edition of the JCS is devoted to personhood, a key issue which certainly deserves the attention. I'm afraid (or perhaps we should be glad?) there are a number of quite different and possibly incompatible perspectives on offer. I thought Eric T Olson's animal problems, expounded in a careful and rather despairing paper ("What are we?", were interesting, but not quite as bad as he thinks.

The issue addressed by the paper, as the title suggests, is that of the nature of human people. Is a person a substance or a process? Persistent over time? Composed of parts, spatially or temporally? Material or non-material? Are we, in fact, anything at all?

One possible stance is that we are simply animals. Probably most of us wouldn't want to say that that is the end of the story, but Olson points out that there does seem to be, as it were, an animal corresponding to each of us. If we're not the animal, and we do the thinking, what's the animal doing? Thinking the same things in pre-established harmony? That seems strange - and if its thoughts are the same as ours, how could we be sure whose thoughts are which? Perhaps the animal doesn't think - but why not: it has a brain and shows lots of signs of mental activity? The idea that the animal might have different thoughts of its own, not corresponding with ours, seems quite bizarre, and difficult to sustain given the way its behaviour matches our thoughts so well.

There's something a little weird about this reasoning. If Olson chose to deny that the animal thinks, he would merely (!) be faced with the problem of why its behaviour matches our thoughts, which appears to be just a somewhat different slant on the classic mind-body problem as we know and love it.

But he doesn't want to do that, and in fact he goes on to consider a re-application of similar reasoning. Because, after all, for each animal there is also a 'lump of flesh'. The lump of flesh is intimately linked with the animal - if we were so inclined we might say that it constitutes it - but it can't just be identified with the animal. After death the lump might persist in the absence of the animal, or given a partial carnal swap of some kind, the animal might survive longer than this particular lump. So again, then, is it the animal that does the thinking, the lump of flesh, or both?

I think the shift to another level give us the clue to the real nature of the problem. Aren't we really just dealing with the well-established fact that the same thing can have different properties when described in different ways? This isn't a unique feature of people or animals. My car may be referred to as the lump of metal in the drive, my prime means of transport, and my chief contribution to global warming, and it has slightly different properties in each case. But once understood, this sort of thing is not normally felt to be the kind of issue we need to lie awake at night fretting over.

There again, perhaps it's a more bothersome issue when it applies to us. Our life seems to unfold on many different levels, yet consciousness seems undeniably single. Must we plump for

one final way of describing ourselves in ourselves, with the others demoted to secondary status (is the subconscious where the real me is to be found?) or do we somehow have to think that our experienced unity is really a pandemonic chorus?

My personal inclination is to think that personhood resides in the single place where consciousness gives rise to agency, and that the animal (to describe myself in those unflattering terms) has, qua animal, a purely supporting role in that crucial process.

Lost In The Woods

21 July 2007

David Gelernter says that the project of strong artificial intelligence - creating a real, human-style consciousness - is lost in the woods, and that it is highly unlikely, only just short of impossible, that a real mind will ever be put together out of software. Although he believes strongly in the value of less ambitious AI projects, Gelernter has some form as a sceptic, having debated the subject last year with Ray Kurzweil, whose faith in the future of technology is of course legendary.

Though he mentions Jerry Fodor with approval Gelernter appears to be in the main a loyal follower of John Searle so far as philosophy is concerned, quoting the famous Chinese Room thought-experiment and repeating Searle's line that a computer simulation of rain doesn't make you wet. This is very well-trodden ground, where Searle's allies and enemies long ago reached a kind of impasse, and new insights are unlikely to be found: moreover, with that 'almost' impossible, Gelernter actually pulls the punch in a way that Searle would certainly never do.

Gelernter suggests that AI has looked towards digital computation for two main reasons; first digital computers seemed in some ways like brains, and second, computation is the leading technology of our day, naturally seen as the best candidate for any theoretically daunting challenge. I think this understates the case a bit. Historically, as far back as Babbage and Turing, the creators of digital computers didn't just design their machines and then notice a resemblance to human brains; they set out to make machines that did what brains do. It's not unreasonable that machines designed in imitation of brains should be what we turn to when we want an artificial mind. Moreover, it's not just that digital computers are the cutting edge of technology at the moment; they have an open-ended flexibility - they are universal machines after all - which is far more brain-like than any other artefact we can imagine. So while it may be right or wrong for AI practitioners to look towards digital computation, there are pretty good reasons for them to do so.

Gelernter's main proposition, in any case, concerns what he calls the cognitive continuum. Consciousness is not just on or off, he observes: it may be closely focussed on a specific object, or it may be in a freer, looser state, in which the mind is more prone to wander. These looser states are not just the brain working inefficiently: they exhibit the mind's ability to associate freely, think emotively, and come up with analogies. These kinds of thinking, especially analogical thinking, are crucial to the nature and success of human consciousness and they deserve more attention within AI. Gelernter himself describes these as only pre-theoretical ideas, but I think he's certainly touched on an interesting area. There's no doubt he's right about consciousness having fuzzy edges, as it were. It would be great to have the range of possible different conscious states properly clarified, but I think that would involve more than a single spectrum - there are surely several different variables at work. To quote just a few examples, besides being focussed or diffuse, our thoughts may be explicit or implicit; they may operate in several different representational modes (thinking in pictures, in words, or neither, for example); they may be accompanied by second-order thoughts (awareness of our awareness) or not (though some would dispute that); they may be under our deliberate direction, roaming

free, or directed by events in the world. They may even be operating in two or more different ways at once, as when we mentally plan a meeting while driving along a busy road.

Gelernter's idea seems to be that when we are concentrating on something, our ideas are operating under close control and they do as they're told: when our minds are wandering, the ideas can collide with each other more or less at random and produce fruitful results. I think he's right to stress the importance of analogical thought, but I'm not altogether sure that it is uniquely enabled by a loose focus. Sometimes when we're thinking hard about a specific problem we may cast about for ideas quite deliberately or look for telling analogies. The cases where a good idea comes into our mind when we're thinking of something else, or of nothing, actually strike us as remarkable, and that may be why we remember and emphasise them more than the cases where a good analogy was put together by careful and conscious hard work.But even if we suppose that a relaxed state of mind is conducive to analogising, that is certainly only part of the story, perhaps the less interesting part. Good analogies don't come from the random combination of ideas - that would be a hopelessly inefficient way of generating useful thoughts - they embody threads of meaning. I suspect that what is going on has something to do with ideas lurking in the mind at a subliminal level - perhaps the relevant neurons are firing too slowly for the thought to be conscious, but fast enough to predispose the mind towards a particular thought. When a related idea, or latent idea, comes along, it is enough to tip a fully-formed new thought into consciousness. I'm getting a bit pre-theoretical here myself, but at least I think it's clear that the old problem of relevance has something to do with this.In fact see an analogy (ha) with iconic representation.

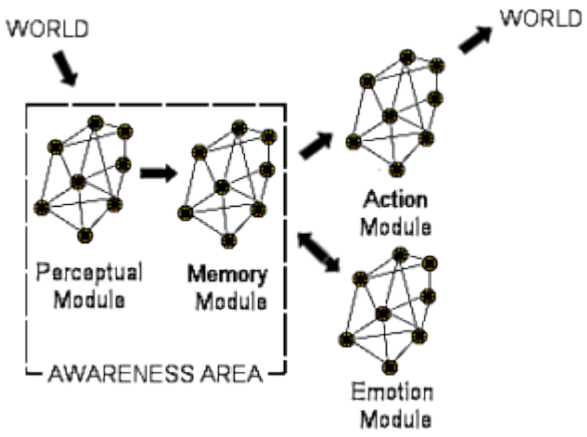
One of the many proposed routes for constructing a theory of meaningfulness is that it starts with simple iconic representation and builds up from there. If you want someone to think of a man, show them a man, goes the theory: if you can't arrange for a man to be present, do a drawing. The marks on the paper share certain properties of shape and appearance with a real man, and so they call the same thought to mind. Stylise your man a bit, and you have a symbol and before you know where you are you'll be covering your pyramid with votive inscriptions. But it's not as simple as that. If I paint a stick man on the wall, it might mean 'man', but it might mean 'I was here', or 'painting' (or of course, 'gentlemen's toilet'). For iconic representation to work, we have to pick out which of the many properties of the icon are relevant: humans are very good at this, but digital computers can't really do it at all. In the same way, a good analogy links two things in different realms whose relevant properties are the same while their other properties are entirely different.

You would perhaps expect Gelernter to make this analogical thinking another basis for his scepticism about strong AI, but in fact he believes the process could and should be simulated by computer, giving us a new generation of vastly more useful artificial intelligence working to principles much more like those of the human brain. It would look a lot more like human consciousness but, he says, it would not be real - just a simulation. I think he is bound to run into new varieties of the old problems with relevance which in various shapes have dogged the progress of AI for many years, but it's interesting to contemplate the strange situation which would arise if he succeeded. Hypothetically, Gelernter's machine would be talking and behaving like a conscious human: many, I dare say, would be happy to accept that it had consciousness of some sort - but not Gelernter.

Axioms of Consciousness

5 August 2007


I've been meaning to mention the paper in Issue 7 of the current volume of the JCS which had about the most exciting abstract that I remember seeing. The paper is "Why Axiomatic Models of Being Conscious?" . Abstracts of academic papers don't tend to be all that thrillingly written, but this one, by Igor Aleksander and Helen Morton, proposes to break consciousness down into five components and offers a 'kernel architecture' to put them back together again. It offers 'an appropriate way of doing science on a first-person phenomenon.



The tone was quite matter-of-fact, of course, but merely analysing consciousness to this extent seems remarkable to me. The mere phrase 'axioms of consciousness' has the same kind of exotic ring as 'the philosopher's stone' in my mind: never mind proposing an overall architecture. Is this a breakthrough at last? Even if it's not right, it must be interesting.


Yes, my eyebrows practically flew off the top of my forehead at some of the implied claims in that abstract. But not altogether in a good way. Perhaps we really are in the same realm of myth and confusion as the philosopher's stone. Any excitement was undercut by the certainty that the paper would prove to have missed the point or otherwise failed to deliver; we've been here so many times before. Moreover, you know, even those axioms are not exactly a new discovery - Aleksander, with a different collaborator, first floated them in 2003.


OK, so maybe I missed them at the time, but surely we ought to look at them with an open mind rather than assuming failure is a certainty? The five are stated in the first person, in accordance with the introspective basis of the paper.¶

- Presence: I feel that I am an entity in the world that is outside of me
- Imagination: I can recall previous sensory experience as a more or less degraded version of that experience. Driven by language, I can imagine experiences I never had.
- Attention: I am selectively conscious of the world outside of me and can select sensory events I wish to imagine.
- Volition: I can imagine the results of taking actions and select an action I wish to take.
- Emotion: I evaluate events and the expected results of actions according to criteria usually called emotions.

That seems an interesting and pretty comprehensive list, though clearly it's impossible to adopt a bold approach like this without raising a lot of different issues.

 : You can say that again. Just look at number 5, for example. Apparently emotions are criteria? My criteria were so strong I just burst into tears? The sight of my true love's face filled my heart with an inexpressibly deep criterion? And their role is to help evaluate events and the result of actions? I mean, apart from the fact that emotions generally interfere with our evaluation of events and actions, that just isn't the essence of emotion at all. I can sit here listening to Bach and be swept along by a whole range of profound emotions which I can hardly even put a name to - I feel exalted, energised, but that doesn't really cover it - without any connection to events or possible actions whatever.

If that isn't enough, Aleksander and Morton claim to be developing an introspective analysis, rather than a functional one. But emotions, it turns out, are basically there to condition our actions. So that's an introspective definition?

 Not the point. We're not trying to capture the ineffable innerness of things here, we're trying to set up a scientific framework; and from the vantage point of that framework - guess what? It turns out the ineffable innerness is entirely negligible and adds nothing to our scientific understanding.

I said in the first place there are two parts to the enterprise: the analysis and then the construction of the architecture. The analysis starts from an introspective view, but it would be absurd to think an architecture could have no regard to function. If that pollutes the non-functionalist purity of the approach from a philosophical point of view, who cares? We're not interested in which scholastic labels to apply to the theory, we're interested in whether it's true.

 Well, you say that, but Aleksander and Morton seem to want to draw some philosophical conclusions - and so do you. They reckon their analysis shows that there is no real 'hard problem' of consciousness. I'm not convinced. For one thing, their attack seems to be quite tightly tied to the Chalmersian version of the hard problem. If subjective states can't be rigorously related to physical states, they say, it opens the way to zombies, people who are physically like us but have no inner life. That would be Chalmers' view, maybe, and I grant that it has attained something close to canonical status. But it's quite possible to disbelieve in the possibility of zombies and still find the relation between the physical and the subjective profoundly problematic.

Much more fundamental than that, though, their whole approach seems to me yet another instance of a phenomenon we might call 'scientist's slide'. Virtually all the terms in this field can be given two values. We can talk about a robot's 'actions', meaning just the way it moves, or we can talk about 'actions', meaning the freely-willed deliberate behaviour of a conscious agent. We can talk about our PC 'thinking' in an inoffensive sense that just means computational processing, or we can talk about 'thinking' in the subjective, conscious sense that no-one would attribute to an ordinary computer on their desk. To put it more technically, the same words in ordinary language can often be used to refer either to access consciousness, a-consciousness, the 'easy problem' kind, or to phenomenal consciousness, p-consciousness, the 'hard problem kind'.

Now over and over again scientists in this area have fallen into the trap of announcing that their theory explains 'real' p-consciousness, but sliding into explaining a-consciousness without

even noticing. Not that the explanation of a-consciousness is trivial or uninteresting: but it ain't the hard problem. I suspect it happens in part because people with a strong scientific background have to overcome a lot of ingrained empiricist mental habits just to grasp the idea of p-consciousness at all: they can do it, but when they get involved in developing a theory and their attention is divided, it just slips away from them again. Not meaning to be rude about scientists: it's the same with philosophers when they just about manage to get their heads round quantum physics, but in discussion it transmutes into magic pixie dust.

 No, no: you're mistaking a deliberate assertion for a confusion. It's not that Aleksander and Morton don't grasp the concept of p-consciousness: they understand it perfectly, but they challenge its utility.

Let me challenge you on this. Take a case where we can't shelter behind philosophical obfuscation, where we have to make a practical decision. Suppose we talk about the morality of killing or hurting animals. Now I believe you would say that we should not harm animals because they feel pain in somewhat the same way we do. But how do we know? Philosophically you can't have any certainty that other humans feel pain, let alone animals.

In practice, I submit, you base your decision on knowledge of the nervous system of the creature in question. We know human beings have brains and nerves just like ours, so we attribute similar feelings to them. Mammals and other creatures with large brains are also assumed to have at least some feelings. By the time we get down to ants, we don't really care: ants are capable of very complex behaviour, maybe, but they don't have very big brains, so we don't worry about their feelings. Plants are living things, but have no nervous system at all, and so we care no more about them than about inanimate objects.

Now I don't think you can deny any of that - so if we stick to practical reasoning, what are we going to look at when we decide the criteria for 'proper' consciousness? It has to be the architecture of the brain and its processes. That's where the paper is aiming: rational criteria for deciding such issues as whether an animal is conscious or whether it has higher order thought. If we can get this tied down properly to the relevant neurology, it may, you know, be possible to bypass the philosophy altogether.

Teacups and Mirrors

17 August 2007

Mirror neurons have been widely described as a crucial discovery and possibly 'the next big thing' (I'm not sure, when I come to think about it, what the last big thing was). Ramachandran describes them as 'empathy neurons' or even 'Dalai Lama neurons', and others have been almost equally enthusiastic. But are they really so good? The trenchant title of a short paper by Emma Borg asks 'If mirror neurons are the answer, what was the question?'.

Your mirror neurons fire both when you perform an action, and when you see someone else perform that action. Borg contrasts them with 'canonical neurons', which fire in response to an object offering the right kind of affordances. In other words, if I've got it right, we have a large group of neurons that fire when we, for example, take a sip of tea: some of them are mirror neurons which also fire when we see someone else drink; others are canonical neurons which also fire when we see a cup or a teapot - 'tea-drinking things'.

At a basic level, the argument that mirror neurons might help to explain empathy, or our understanding of other people, is clear enough. When I see A do x, the mirror neurons mean my mental activity has at least some limited features in common with A's (presumed) mental activity, or at least what A's mental activity would be if A were me. You can see why this resembles telepathy of a sort, and it seems a natural hypothesis that it might form the basis of our understanding of other people. One of the many theories on offer to explain autism, in fact, holds that it is caused by a deficiency in mirror neuron activity. Apparently there is evidence to show that autistic people don't show the same kinds of activity in the relevant regions as normal people when they observe other people's behaviour. It could be that the absence of mirror neuron activity has left them with no basis for a 'theory of mind': of course it could also be that the absence of an effective theory of mind, caused by something else altogether, is somehow suppressing the activity of their mirror neurons.

Borg's target is the idea that mirror neurons in themselves give us the ability to attribute high-level intentions to other people, by running simulated intentions of our own that match the observed actions of the other person. The initial idea is roughly that when we see someone lift a cup, some of our neurons start doing that tea-cup lifting thing in sympathy (off-line in some way, of course, or we should grab a cup ourselves). This is like harbouring the intention of lifting the cup, but we are able to attribute the intention to the other person. However, this only gets us as far as the deliberate lifting of the cup: it has been further claimed that mirror neurons give us the ability to deduce the over-arching intention - drinking a cup of tea. The claim is that mirror neurons not only resonate with the current action but also more faintly (or rather, in smaller numbers) with the next likely action, and this provides a guide to the higher-level activity of which the single act is part.

Borg points out that actions in themselves are highly ambiguous. I may lift a cup to test its weight, or to stop you getting it, rather than in order to drink from it. It's certainly not the case

that every basic act dictates its successors, or we should be trapped in a cycle of stereotyped behaviour. When we run our mental simulation then, how can we know which secondary echoes we need to start off in our mirror neurons - unless we already know which higher-level course of action we are dealing with? In short, mirror neurons are not enough unless we already have a working theory of mind from some other source.

We might argue that we don't need to know the intention in advance, because the simulation allows us to test out several different higher-level courses of action at once. But again, the mere observation of the single act before us won't allow us to choose between them. In the end we'll always be driven back to appealing to something more than mere mirror neuron activity. None of this suggests that mirror neurons are uninteresting, but perhaps they are not, after all, going to be our Rosetta Stone in deciphering the brain.

Borg describes her argument as anti-behaviourist, resisting the idea that intentions and other 'mentalistic' states can be reduced to simple patterns of activity. Fair enough, but given that behaviourism doesn't put up much of a fight these days, it may be more interesting that it bears a distinct resemblance - or so it seems to me - to many other problems which have afflicted attempts to reduce or naturalise intentionality, up to and including the frame problem. It's as though we were trying to find a way through an impenetrable hedge: every so often someone finds a promising looking thin patch and starts to shove through; but sooner or later they meet one or another stretch of suspiciously-similar looking brick wall.

Benjamin Libet

8 September 2007

I've just heard that Benjamin Libet died at the end of July.

His famous experiments remain surprising and controversial, and in fact I remember discounting them altogether when I first heard about them. It must have been in about 1985, I think: we were sitting on rickety chairs in the trademark squalor of Gordons Wine Bar, and I was giving the company the benefit of what I supposed were my more sophisticated views on the subject of free will.

"That's all very well," said a young scientist, "But there's this bloke who's proved experimentally that free will does not exist... Libet, I think. I was reading about it last week."

"That sounds interesting," I said, "But I don't think that's possible in principle. Freedom is not an observable physical property, so the issue is beyond the reach of empirical methods. Perhaps you're thinking that free will requires some kind of causal discontinuity, but that isn't actually the case."

"Well, these experiments show that the decision to act is made before we become aware of it. That proves our conscious thoughts about the decision don't determine which way it goes."

"I don't really see how you could determine experimentally when a decision is made, other than by asking the experimental subject," I said, "It's not as if you can read off the contents of someone's mind from an encephalogram."

"I'll get you the reference." he said - alas he never did, and it was some years before I got round to mitigating my ignorance of Libet's experiments. Libet himself seems to have been a thoughtful, complex man, very far from the gung-ho debunker I had at first envisaged.

Whatever the fate of Libet's theory about a conscious mental field, his experiments are surely classics in the field and guarantee him a permanent place of honour in its history.

Joycean Consciousness

16 September 2007

I've been reading David Lodge's 'Consciousness and the Novel'. Lodge points out that novelists share with cognitive scientists the aim of exploring human consciousness but have characteristically different methods and concerns; typically concentrating on the idiosyncratic, qualia-ridden details of individual experience rather than seeking to extract generalisable scientific laws.

Lodge is good at technical analysis of the means by which novelists obtain their effects (I'm surprised he manages to write readable fiction while carrying such a weighty theoretical apparatus in his head - I should have thought it would induce a paralysing self-consciousness). In the first chapter he presents the development of narrative styles as being in part a search for more effective ways of conveying the reality of human conscious experience. In particular, he interprets modernist and stream-of-consciousness novels as attempting to break through the stale conventions of the traditional novel and give an altogether more immediate impression of consciousness from the inside. James Joyce is in his eyes the most radical and successful of the authors to make this attempt - in fact

'He came as close to representing the phenomenon of consciousness as perhaps any writer has ever done in the history of literature.'

Well, up to a point. I think the Joycean style (as exemplified in Ulysses) is simply based on a new and different literary convention which is in reality at least as far from ordinary consciousness as the prose of Jane Austen. That explains why Joyce is, frankly, hard going at times and remains the preserve of the intellectual reader. If it were a vivid unmediated gateway into the phenomenal experience of the characters, Joycean prose would be the easiest to write and read and he would surely be among the most popular and accessible of authors.

This sort of paradoxical development - the attempt at nature which produces new artifice - is far from unique in Eng Lit, of course: it's hard for modern readers to believe that the Lyrical Ballads, for example, could be presented as a breakthrough in the use of everyday language shorn of poetic affectation, yet start with something as peculiar as The Rime of the Ancient Mariner. But consideration of what Joyce does and does not achieve is interesting so far as the nature of conscious experience is concerned.

What are in fact the contents of consciousness? Drawing on introspection I find that some of them are indeed properly formed verbal sentences: as I sit here writing, fragments of text are replayed through my mind, adjusted and echoed immediately before being typed. Explicit words feature in my consciousness on other occasions too, but these occasions seem to me the exceptions and full-blown English sentences are certainly not the chief medium of my conscious experience.

That chief medium seems to be composed of thoughts which are as clearly focussed and meaningful as pieces of text, but have none of the same structure. I may think that a noise outside is probably the postman: to think so takes no appreciable time and does not involve the

rehearsal of a formula over one or two seconds, as verbal thought might do. At the same time, I can go on thinking the same thought for a while if I choose, and when I think something else the transition may be abrupt, but is often smooth and unclear.

This is partly because, beside the mainstream thoughts there are a number of other candidate thoughts lurking in the penumbra of consciousness. I may be thinking about what I'm writing, but I'm sort of aware that I might think about whether it's time to leave quite soon. A transition between one thought and another may therefore be a matter of one coming forward and another receding - though it can be the sudden intrusion of an altogether new thought, too. Although this penumbra business is undoubtedly vague, I think it is always clear that only one thought is actually being thought at any given time.

At the risk of schematising too much, we could say that the mainstream of thought is accompanied by an intermittent superstructure of explicit verbal thoughts and often by a hazy underground of potential thoughts. But the mainstream itself is fluent and non-verbal, though just as meaningful as words. All my thoughts, moreover, come with a kind of background. When I think that a sound outside might be the postman, I do not have to think about the history of the Post Office, my uncle who once worked for it, or the items of post I am expecting some time soon; but all those items are somehow readily available; not in quite the same way as the candidate thoughts I mentioned above - they're not sort of wanting to be thought about, they're just available.

Besides this, there is a general awareness of the world around me. Sometimes this recedes into a penumbra like the one in which the candidate thoughts lurk - I'm not quite sure whether it is in fact the same penumbra, nor whether the spotlight of attention is the same for thoughts and perceptions. I rather think I can pay attention to a perception and a thought at the same time, but it may be that they have to jostle for essentially the same attention, and at best merely alternate at the forefront of my mind. I'm not quite sure.

The objects I do perceive all come with a ready-made interpretation. When I look at the screen, I see the text of a blog post, but if I wish I can change the interpretation and see characters in a particular font, arrays of pixels, or my computer. Again, I don't quite know whether these interpretations are the same kind of thing as the background awareness that comes with my thoughts. There seems to be a similarity, but the interpretations are required in a way that the background isn't. I can't look at the screen without giving it some interpretation, but I can think about the postman while ignoring all the associations that he brings with him.

A special case of my awareness of the world is my awareness of my own body, which I think exceptionally does not require interpretation (there's only one thing a pain in my foot can be), but which is closely associated with my awareness of my own emotions. Emotions may be linked with a racing heart or a tightness in the stomach, but they are also linked with the most complex and sophisticated of my thoughts. I have a general emotional background (fairly calm at the moment, but tinged with some slight tension about when I need to leave the house, and the prospect of seeing the dentist later), and I also have emotional reactions to particular perceptions and thoughts.

I have certainly omitted and misrepresented a good deal in the foregoing, but I hope this sketch of the contents of consciousness is recognisable enough.

Novelists of any school clearly face some severe problems. One is the inherent indescribability of the qualia which make up or accompany our perceptions, and of the inner nature of

emotions. Two others appear equally insoluble. One is how to represent thoughts which are meaningful, but not verbal in form. One way is to translate the thoughts into words, and present them as if they were like utterances of the character in question: but this short-changes us on the fluid and instantaneous nature of thought, and breaks the link with the background; it also destroys the difference between mainstream thoughts and thoughts in actual words. The alternative is to describe the content of the thought, but this loses the sense of vivid intentionality which thoughts have and condemns us to a third-person view, outside the consciousness being depicted.

The other insoluble problem is how to present the simultaneous complexity of consciousness when the structure of prose obliges us to relate one thing at a time in sequence.

Traditional novelists use a mixture of translation and description to depict a character's thoughts, adjusting the balance to suit their own purposes. Lodge says the ultimate development in this direction is free indirect style, in which the thoughts are translated into words but not presented as quotation. He regards Jane Austen as supreme in the use of this style; it is a particularly flexible and fluid approach which helps the author get the best of both translation and description while remaining lively and vivid.

I believe the second problem is mainly addressed in traditional novels by the use of carefully chosen implication, and by the even more powerful tool of implicature, where something omitted or added from the account indicates to the reader a large amount of subsidiary information which is not explicitly given in the text. This approach allows the author. to some extent at least, to evoke some of the background to the character's thoughts and also to indicate more than one level of consciousness at once.

However, although these strategies work quite well in traditional novels, there is a degree of artifice about them which may detract from the reality of the portrayal, and distance the reader from the consciousness of the character. How much better, then, to simply transcribe the stream of consciousness as it actually occurs in all its confusing heterogeneity? The trouble is that the two problems touched on above are still there. What Joyce tends to give us, in the famous last chapter of Ulysses, for example, is a straight translation of the mainstream of thoughts. Sentences and other grammatical structures are ignored or broken down to help indicate the rapidity and fluidity of the mainstream, but in essence we only get one level of consciousness. Since the author has undertaken to give us everything at that level he can no longer build in helpful implicatures by adding or leaving out carefully chosen items: the reader has to cope with the whole flood of thoughts, and is forced to do more work than normal to try to infer the background. An unexpected side effect is that not much emotional tone comes through: the words seem rather like the affectless, breathless drone of someone in a trance or in a state of delirium. Joyce succeeds in making it readable by actually practising some covert selection of items and moderating the grasshopper leaps and repetitions which I think would occur in a more rigidly realistic transcription of mainstream consciousness.

Does it work, on the whole? I don't think it delivers what was promised. In 1922 Joyce said:

"In Ulysses I have recorded, simultaneously, what a man says, sees, thinks and what such seeing, thinking, saying does to what you Freudians call the unconscious..."

which I think pretty much amounts to a claim to have delivered all the contents of consciousness sketched out above. I don't think Joyce really did this, or if he did, it was by the

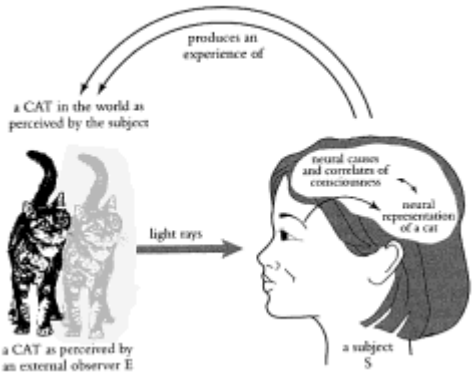
crafty use of traditional novelistic skills within a modernist presentation. But that doesn't mean Ulysses isn't interesting and a considerable achievement.

I think the right way to see it is in the context of a general resort to unconventional means in the arts which took place in the nineteenth and twentieth centuries. By the middle of the nineteenth century or thereabouts, the technical problems of painting, music and writing had pretty much been solved to the extent that they ever could be, and optimum conventions and methods had been established. Our culture being what it is, it was impossible for creative people to sit back and spend a few centuries exploiting these achievements: they had to press on and deliver something new. Since there were no more problems to solve, this novelty had to come from approaches that were radically different rather than improvements on what had gone before: various forms of unrealism and abstraction, twelve-tone or atonal music - and the stream of consciousness in the novel. These novel approaches tended not to lead on to anything - Joyce has no real successors - nor to achieve great popularity, but they did at least open up new imaginative and creative possibilities.

So far as consciousness is concerned, we can perhaps see Joyce's work as a bold attempt to offer an analysis of consciousness as a single, transcribable phenomenon: in my eyes it proves the opposite: that consciousness is inherently complex and no single simple theory can capture the whole thing.

Reflexive Monism

6 October 2007



Max Velmans has produced Reflexive Monism as a valiantly renewed effort to sort out the confused story of the relations between observer, object, and experience. This is one of those subjects that really ought to be perfectly straightforward but in fact has descended into such a dense thicket of philosophical clarification that the scope for misunderstanding and talking past one another is huge. In my more pessimistic moments I wonder whether the issue is really reclaimable at all, at least by means of further discussion.

Velmans' view apparently stems from a revelation he experienced when he noticed that the world he had heretofore regarded as the public, objective, external one – the world we all experience, full of cats and a number of other things – was in fact a phenomenal world, a world as experienced by him. Those things out there are our experiences of the world, and so reflexive monism belongs with the externalist theories that seem to have become popular recently. It is also a dual aspect theory; that is, the one underlying stuff in which all monist s must believe expresses itself in two ways, as the objective physical world and as the consciously experienced phenomenal world we actually see out there.

Dual aspect theories seem attractively sensible, and very probably true as far as they go; but I think one can be pardoned for still feeling slightly unsatisfied by them. Okay, so the world doesn't consist of two kinds of stuff, it just has two aspects; but to round that out into a proper explanation we need an account of what an aspect might be. Ideally we also want an account of the nature of the one underlying stuff which would explain why on earth it expresses itself in two different ways. These accounts are not easy to give: in fact it is quite difficult to say anything at all about the fundamental stuff, much as all good metaphysicians must wish to do so. I don't think these issues are quite so problematic for the poor benighted dualists, who have a good reason why things might appear in two different guises, or the more brutal kinds of monist, who can as it were, just tick all the 'No' boxes on the form. Velmans, in fairness, has given us a helpful hint in the name of his theory: the reflexivity of his monism refers to the idea of the Universe becoming aware of itself through the medium of conscious individuals, which at least suggests where the two aspects might spring from.

Velmans brings out well what I think is the main attraction of externalism: that it eliminates the idea that the objects of perception are entirely in the head, that all we ever really experience are representations in the brain. Some of Velmans' assaults on this idea, however, are a little dubious. Take his 'skull' argument: the phenomenal world extends as far as we can see – to the horizon and the dome of the sky, he suggests. Now if the phenomenal world is all inside my brain, my real, non-phenomenal skull exists beyond that world: beyond the dome of the sky. How ridiculous is that? Rhetorically, this conjures up the idea of the real skull floating in outer space, or perhaps enclosing the world like a second, bony sky. But really, in saying that the real skull is beyond the phenomenal world, we don't mean it's geographically or spatially a bit beyond it: we mean it's in another world altogether; in a different mode of existence.

There's a similarly questionable treatment of location in Velmans' references to those materialists who would see experience as constituted by functions or patterns of neuron firing in the brain. Velmans again attributes to such people the belief that experience is actually located in the brain. They might well agree, but perhaps with some reservation: I think many or most functionalists, for example, would distinguish between a particular instantiation of a function, which certainly exists in a physical place in the brain, and the function itself, which could be run by other brains, and which exists in some Platonic realm where spatial position is irrelevant or meaningless.

The main problem for me, as with some other versions of externalism, is whether reflexive monism delivers the simplification it seems to promise or merely relocates the problem. Velmans provides diagrams which illustrate the difference between a dualist view, where the phenomenal perception floats above observer and object in a mental/spiritual world, a reductionist view where the percept is similarly dangling, and his own, where we have the object (a cat, as it happens), the observer, and nothing more than a couple of arrows. The trouble is, we still actually have the real cat and the perceived, phenomenal one: Velmans has pulled off a sly bit of legerdemain and shuffled one under the other.

Velmans spends some time expounding the idea of 'perceptual projection' – that the phenomenal world is projected out there into physical space – and defending himself against the charge of smearing real cats with phenomenal cat-perception stuff; but I think there is a worse difficulty. The phenomenal experience may have been projected out of our skulls, but it's still all we get to deal with, and that seems to leave us dangerously isolated, close to the beginning of the broad and easy downward path which leads to solipsism. It's not so much that the danger is inescapable – more that I'm left wondering whether taking all those phenomenal experiences out into the external world actually changed things all that much.

Velmans wraps up with an exposition of how his view impinges on the hard problem. In essence, he thinks that when we've grasped reflexive monism properly, we will see that the fact that the world has two different aspects is just one of those features which, although they are slightly mysterious, there is no need to worry about. We don't agonise, he suggests, over the "hard problems" of physics – why does an electric current in a wire give rise to a magnetic field? Why do electrons behave like waves in some circumstances and like particles in others? Why is there any matter in the Universe at all?

Actually, I think people do agonise over those problems, as it happens. I must have spent a considerable number of shortish and frustrating periods of time wondering in vain why there was anything.

Anthropomorphism

17 October 2007

I came across *A defense of Anthropomorphism: comparing Coetzee and Gowdy* by Onno Oerlemans the other day. The main drift of the paper is literary: it compares the realistic approach to dog sentiment in J.M.Coetzee's Disgrace with the strongly anthropomorphic elephants in Barbara Gowdy's The White Bone. But it begins with a wide-ranging survey of anthropomorphism, the attribution of human-like qualities to entities that don't actually have them. It mentions that the origin of the term has to do with representing God as human (a downgrade, unlike other cases of anthropomorphism), notes Darwin's willingness to attribute similar emotions to humans and animals, and summarises Derrida's essay The Animal That Therefore I Am (More To Follow). The Derrida piece discusses the embarrassment a human being may feel about being naked in front of a pet cat (I didn't know that the popular Ceiling Cat internet meme had such serious intellectual roots) and concludes that taking the consciousness of animals as seriously as our own threatens one of the fundamental distinctions installed in the foundations of our conception of the world.

That may be, but the attribution of human sentience to animals is rife in popular culture, especially when it comes to children. Some lists of banned books suggest that the Chinese government has cracked down on many apparently harmless children's books; it turns out this is because at one time the Chinese decided to eliminate anthropomorphism from children's literature, wiping out a large swathe of traditional and Western stories. I can't help feeling a small degree of sympathy with this: a look at children's television reveals so many characters who are either outright talking animals, or (even stranger) humanoids with animal heads, you might well conclude there was some law against the depiction of human beings. It would surely seem odd to any aliens who might be watching that we were so obsessed with the fantasised doings of other species.

Or perhaps it wouldn't seem strange at all, and they would merely make plans for a picture-book series about Hubert the Human, his body spherical with twelve tentacles just like a normal person, but his head displaying the strange features of Homo Sapiens. It seems likely that our fondness for anthropomorphism has something to do with our marked tendency to see faces in random patterns: our brains are clearly set up with a strong prejudice towards recognising people, or sentience at least, even where it doesn't really exist. Such a strong tendency must surely have evolved because of a high survival value - it seems plausible that erring on the side of caution when spotting potential enemies or predators, for example, might be a good strategy - and if that's the case we might expect any aliens to have evolved a similar bias.

That bias is a problem for us when it comes to science, however. When considering animal behaviour, it seems natural, almost unavoidable, to assume that the same kind of feelings, intentions and plans are at work as those responsible for similar behaviour in humans. After all, humans are animals. It's clear that other animals don't make such complicated plans as we do; they don't talk in the same way we do and don't seem to have the same kinds of abstract thought. But some of them seem to have at least the beginnings or precursors of human-style consciousness.

Unfortunately, careful observation shows beyond doubt that some forms of animal behaviour which seemed purposeful are really just fantastically well-developed instincts. The sphex wasp seems to check its burrow out of forethought before dragging its victim inside; but if you move the victim slightly and make it start the pattern again, it will check the burrow again, and go on doing so tirelessly over and over, in spite of the fact that it knows, or should know, that nothing could possibly have gone into the burrow since the last check.

A parsimonious approach seems called for. The methodologically correct principle to apply was eventually crystallised in Morgan's Canon:

'In no case may we interpret an action as the outcome of the exercise of a higher mental faculty, if it can be interpreted as the exercise of one which stands lower in the psychological scale'

In effect, this principle sets up a strong barrier against anthropomorphism: we may only attribute human-style conscious thought to an animal if nothing else - no combination of instinct, training, environment and luck - can possibly account for its behaviour. I said this was 'methodologically correct', but in fact it is a very strong restriction, and it could be argued that if it were rigorously applied, the attribution of human-style cognition to certain humans might begin to look doubtful. According to Oerlemans, ethologists have been asking whether, by striving too hard to avoid anthropomorphism, we haven't sometimes denied ourselves legitimate and valuable insights.

It's interesting to reflect that in another part of the forest we are faced with a similar difficulty and have adopted different principles. Besides the question of when to attribute intelligence to animals, we have the question of when to do so for machines. The nearest thing we have to Morgan's Canon here is the Turing Test, which says that if something seems like a conscious intelligence after ten minutes or so of conversation, we might as well assume that that's what it is. Now as it happens, because of its linguistic bias, the Turing Test would not admit any animal species to the human level of consciousness; but it does seem to be a less demanding criterion. Perhaps this is because of the differing history in the two fields: we've always been surrounded by animals whose behaviour was intelligent in some degree, and perhaps need to rein in our optimism; whereas there were few machines until the nineteenth century, and the conviction that they could in principle be intelligent in any sense took time to gain acceptance - so a more encouraging test seems right.

Perhaps, if some future genius comes up with the definitive test for consciousness, it will lie somewhere between Morgan and Turing, and be equally applicable to animals and machines?

Philosophy and Neuroscience

28 October 2007

I've just caught up a bit belatedly with The Philosophy and Neuroscience Movement, the paper of Pete Mandik's, written with Andrew Brook, which he featured on his blog at the beginning of the month. It's an interesting read and seems to have garnered quite a bit of attention.

The general question of relations between science and philosophy has of course been a vexed one ever since the distinction started to be made (I don't think someone like Isaac Newton would have seen any particular gulf between the two). Some scientists speak of philosophers the way an aristocrat might talk about the gypsies encamped on the edge of his land: with patronising disapproval, but also with a slight secret fear that they might have obscure magic secrets which could ruin the scientific harvest: philosophers for their part have been largely scared away from the grand metaphysical uplands by the obvious dangers in ontologising without a firm grasp of modern physics.

When it comes to philosophy of mind and neuroscience the intertwining of the issues creates an especially great need for co-operation and interaction, and perhaps holds out some prospect of more of a meeting on equal terms. Some thinking along these lines must, at any rate, have lain behind the movement for working together which Brook and Mandik say began about twenty-five years ago; they rightly point out that the influence of scientific advances has transformed the way we think about language, memory and vision (in fact scientific understanding of the visual system has been influencing philosophical ideas for hundreds of years: the discovery of images on the retina must surely have predisposed philosophers towards thinking in terms of an homunculus, a 'little man' sitting and watching the pictures, and perhaps still adds some weight to the perceived importance of mental images as key intermediaries between us and reality.

But I think it's still not uncommon for people on either side to assume that they can get on quite well on their own. Neuroscientists may be tempted to think they should just get on with the science, and let the philosophy take care of itself: maybe the philosophical answers will come along a s a kind of bonus with the scientific results - and if they don't, well philosophers never answer anything anyway, do they? Philosophers, equally, may suppose the science is a matter of detail - oh, by the way, they tell me that the firing of c-fibres is not actually equivalent to pain, as it turns out, but it doesn't really matter if it's, you know, f-fibres or z-fibres: for the sake of brevity in this discussion let's just pretend it is c-fibres...

Hence in part, I suppose, the paper, which discusses (necessarily in brief and summary form) a number of interesting areas of interaction which show how much there is to be gained by positive engagement.

An interesting one is the suggestion that neuroscience and philosophers of neuroscience have greatly strengthened the case against nomological (law-based) theories of scientific explanation. Physics tends to be taken as the paradigmatic science, but one of its leading

characteristics is its amenability to nomological treatment. Physics produces clear universal laws which work with precise mathematical accuracy. In biology, things are very different, and we tend to have to work with explanations which are teleological (concerned with the purpose of things) or statistical in nature. This appears to be especially true in neuroscience, where the brain exhibits complex structure but does not seem to operate under simple laws.

Another area I found thought-provoking was the claim, which I found surprising at first, that new neuroscientific tools have led to a renaissance in introspective studies. Introspection, the direct inward examination of the contents of one's own consciousness, was a no-go area for many years after the collapse of attempts to systematise introspectionist findings (indeed, we've been reading recently in the JCS about the disgust with introspectionism that led J.B. Watson to set up as a radical behaviourist, declaring that there actually were no damn contents of consciousness). The trouble with introspection has always been that the results are chaotic and unconfirmable: training designed to improve results raises the new danger of bias and getting out of your subjects only what you trained into them. How could this situation possibly be reclaimed?

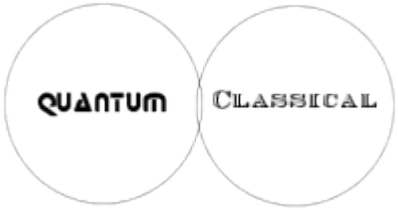
Yet it's true when you come to think of it that many of the numerous studies with fMRI and other scanners which have taken place in recent years have relied on introspective reports. The thing is, of course, the scanners provide an avenue of confirmation and corroboration which Wundt and Titchener never had and legitimise introspective reports. If it weren't already so securely fastened, this would be another nail in the coffin of radical behaviourism.

A third point among many which particularly struck me was the issue of neural semantics. Consideration of the brain and nervous system as information processors takes us into the philosophically intractable area of intentionality. The real difficulties (and the most interesting issues) here are well-known to philosophers but easily overlooked or undervalued by scientists. However, neurobiology offers influential examples of feature-detection mechanisms which might (or might not) point the way to a proper analysis of meaning. My personal view is that this is a particularly promising area for future development, where ancient mysteries might really be dispelled in part by future research.

The paper concludes with a rather downbeat look at consciousness itself. Faced with claims that consciousness is in part not neural, or even physical, many neuroscientists (and their 'philosophical fellow-travellers') ignore them or 'throw science at it', the authors say. They rightly consider this a risky approach, liable to lead to the familiar syndrome in which the scientists explains a toy-town or simplified conception of consciousness, leaving the really tough problem breathing unacknowledged down their necks. This might be a slightly gloomy view, but it is impossible to disagree when the authors call for better rejoinders to such claims.

Testable quantum effects in the brain?

8 November 2007



The idea that quantum mechanics is part of the explanation of consciousness, in one way or another, is supported by more than one school of thought, some of which have been covered here in the past. Recently MLU was quick to pick up on a new development in the shape of a paper by Efstratios Manousakis which claims that testable predictions based on the application of quantum theory have been borne out by experiment.

The claims made are relatively modest compared to those of Penrose and Hameroff, say. Manousakis is not attempting to explain the fundamental nature of conscious thought, and the quantum theory he invokes isn't new. At its baldest, his paper merely suggests that dominance duration of states in binocular rivalry can be accurately modelled by assuming we're dealing with a small quantum mechanical system embedded in a classical brain - although if that much were true, it would certainly raise further issues.

But let's take it one step at a time. What is binocular rivalry, to begin with? When radically different images are presented to the left and right eye, instead of a blurry mixture of both images, we generally see one fairly clearly (occasionally a composite made of bits of both images); but curiously enough, the perceived image tends to switch from one to the other at apparently random intervals. Although this switching process can be influenced consciously, it happens spontaneously and therefore seems to be a kind of indecisiveness in some unconscious mechanism in our visual system.

Manousakis proposes a state of potential consciousness, with the two possible perceptions hanging in limbo. When the relevant wave function is collapsed through the intervention of another, purely classical brain mechanism, one or other of the appropriate neural correlates of consciousness is actualised and the view through one eye or the other enters actual consciousness. This model clearly requires perception to operate on two levels; one, a system that generates the potential consciousness, and two, another which actualises states of consciousness by checking up every now and then.

Thus far we have no particular reason to believe that the application of quantum concepts is anything more than a rather recondite speculation; but Manousakis has shown that the persistence of one state in the binocular rivalry, and the frequency of switching, can be predicted on the basis of his model: moreover, it explains and accurately predicts another observed phenomenon, namely that if the stimuli in a binocular rivalry experiment are removed and returned at intervals, the frequency of switching is significantly reduced. If I've understood it correctly (and I'm not quite sure I have), one of the two images tends to stay around because when you collapse the wave function a second time, there is a high probability of it collapsing the same way. If you remove the images for a while, no new collapses can occur until the images return. It's as though you were throwing a dice and moving to the other state when you threw a six; if you keep throwing, the changes will happen often, but if you take the dice away for a few minutes between each throw, the changes will become less frequent.

Manousakis has also demonstrated that his theory is consistent with the changes observed in subjects who had taken LSD (a slightly puzzling choice of experiment - I'm slightly surprised it's even legal - but it seems it reflects established earlier findings in the field).

So - a breakthrough? Possibly, but I see some issues which need clearing up. First, to nail the case down, we need better reasons to think that no mere classical explanation could account for the observations. There might yet be easier ways to account for changes in dominance duration. We also need some explanation of why quantum mechanics only seem to apply in unusual cases of ambiguity like binocular rivalry. A traditional theorist who attributes binocular rivalry to problems with the brain's interpretative systems has no trouble explaining why the odder effects only occur when we impose special interpretive challenges by setting up unusual conditions: but if a quantum mechanical system is doing the job Manousakis proposes, I would have thought occasional quantum jumps would have been noticeable in ordinary perception, too.

It would help, in addition, to have a clearer idea of how and why quantum mechanics is supposed to apply here. It could presumably be that potential consciousness is generated at a microscopic level where quantum effects would naturally be observable: an account of how that works would be helpful - are we back with Penrosian microtubules? However, Manousakis seems to leave open the possibility that we're merely dealing with an analogy here, and that the maths he employs just happens to work for both good old-fashioned quantum effects and for a subtle mental mechanism whose basic nature is classical. That would be interesting in a different way.

It will be interesting, in any case, to see whether anyone else picks up on these results.

The Silence of the Apes

10 November 2007

A piece by Clive Wynne reviews the well-known attempts which have been made to teach chimps (or bonobos) to use language and draws the melancholy conclusion that the net result has in the end merely confirmed that grammar is uniquely human. It seems a fair assessment to me (though I always find it difficult not to be convinced by some of the remarkable videos which have been produced) , but it did provoke some thoughts that had never occurred to me in this connection before.

According to Wynne, the chimps show clear signs of recognising a number of nouns, but no sign of either putting the nouns in the right order or recognising the significance of the order in which they have been put by humans. They cannot, in other words, distinguish between 'snake bites dog' and 'dog bites snake', which is a key test of grammatical competence.

But word order is obviously not the whole story so far as grammar is concerned. One of the first things you learn in Latin is that in that language, although there may be a preferred word order, it isn't grammatically decisive: 'Serpens mordet canem' means the same as 'Canem mordet serpens' (to express the reversed relationship, you'd have to say 'Canis mordet serpentem'). Perhaps apes just have trouble with grammars like that of English which rely on word order; perhaps they would do better with a language which used inflection, or some other grammatical mechanism instead? Was failure, in short, built into these experiments just as surely as it was into the doomed earlier attempts to teach them to speak?

Two quite different languages were involved in the different experiments: Washoe and other chimps were taught ASL, a sign language used by deaf humans; Kanzi and others were taught to communicate in specially-created lexigrams, symbols arranged on a keyboard, though the experimenters apparently used spoken English for the most part.

I don't know much about ASL, but it does appear to use word order, albeit a different one from that in normal English; typically the topic is mentioned first, followed a comment. You can do this sort of thing in English of course ('That snake - the dog bit it.'), but it isn't standard. If you want to specify a time in ASL, which might be done with tenses in English, you should mention it first, before the topic. In making your comment, the word order appears to be similar to the standard English one, though there may be some degree of flexibility. My impression is that ASL users would tend to break down the information they're conveying into smaller chunks than would be normal in English, taking a clause at a time to help minimise ambiguity. There is something called inflection in ASL, but it isn't the kind of conjugation and declension we're used to in Latin and doesn't play the same grammatical role. In fact, one important grammatical indicator in ASL is facial expression - a possible problem for the chimps, although they could presumably manage some of the basic head-tilting and eyebrow (alright, brow-ridge) raising.

With lexigrams. the relationship to standard English is closer: each of the 384 lexigrams is equivalent to an English word, and indeed some consist of the word written in a particular shape with particular colours. This obviously makes things easy for the experimenters and in

some ways for the bonobos, who would otherwise be faced with learning two languages, heard English and spoken lexigram. The grammar involved is therefore essentially English, and if anything the use of lexigrams makes word order even more crucial, since verbs are necessarily invariant and there are no plurals: so we don't even get the kind of extra clues we might have in an English sentence like 'The dog bites the snakes'.

Prima facie then, it does seem to me that unless the chimps were naturally at ease with using English-style word order as their sole grammatical tool, they were actually given little scope to demonstrate grammatical abilities by any of these experiments. We can perhaps follow the implications a little further. ASL is not very much like ordinary English in its grammar or structure. The adoption of a different channel of communication by deaf people appears to have called for a very different language. It seems natural to suppose, then, that if we require even more radically different channels to communicate with chimps we need a language even more remote from English. Perhaps both ASL and lexigrams are too strongly adapted for human use: true communication may require a form of language which is novel and as difficult for human beings to learn as the chimps; one in fact which might require some rethinking of how grammar can be expressed (something similar had to happen before it was accepted that ASL and other sign languages had true grammar). But if merely understanding this hypothetical language would be dauntingly difficult for us, it hardly seems probable that we could construct it in the first place.

The only way such a language could be constructed, I think, is if the chimps were able to make an equal contribution from their side, rather than being captives drilled in an essentially human style of communication. If a human and chimp community enjoyed a close but free relationship of real importance to both, possibly based on trade or similar relations of mutual benefit, perhaps the differing conventions of different species could be shared and a kind of pidgin developed, as happened all over the world when Western traders first appeared - although this time it would have to be a non-vocal one. The chances of anything like this happening, if not zero in any case are of course remote, and growing less all the time, so sadly the chances are that if chimps do after all have some grammatical ability, we'll never really know about it.

Laughing Computers

24 November 2007

We've discussed here previously the enigma of what grief is for; but almost equally puzzling is the function of laughter. Apparently laughter is not unique to human beings, although in chimps and other animals the physical symptoms of hilarity do not necessarily resemble the human ones very closely. Without going overboard on evolutionary explanation, it does seem that such a noteworthy piece of behaviour must have some survival value, but it's not easy to see what a series of involuntary and convulsive vocalisations, possibly accompanied by watering eyes and general incapacitation, is doing for us. Shared laughter undoubtedly helps build social solidarity and good feeling, but surely a bit of a hug would be fine for that purpose - what's with the cachinnation?

Igor M. Suslov has a paper out, building on earlier thoughts, which presents an attempt to explain humour and its function. He thinks it would be feasible for a computer to appreciate jokes in the same way as human beings; but the implication of his theory seems to be that a sophisticated computer - certainly one designed to do the kind of thinking humans do - would actually have to laugh.

Suslov's theory draws on the idea (not a new one) that humour arises from the sudden perception of incongruity and the resulting rapid shift of interpretation. When cognitive processes attain a certain level of sophistication, the brain is faced with many circumstances where there are competing interpretations of its sensory input. Is that a bear over there, or just a bush? The brain has to plump for one reading - it can't delay presenting a view to consciousness until further observations have resolved the ambiguity for obvious practical reasons - and it constructs its expectations about the future flow of events on that basis: but it has the capacity to retain one or two competing interpretations in the background just in case. In fact, according to Suslov, it holds a number of branching future possibilities in mind at any one time.

The brain's choice of scenario can only be based on an assessment of probability, so it is inevitably wrong on occasion - hey, it's not a bear, after all! In principle, the brain could wait for the currently assumed scenario to drain away naturally when it reached its current end: but the disadvantages of realising one's error slowly are obvious. Theoretically another alternative would be to delete all recollection of the original mistake: but the best approach seems to be to tolerate the fact that our beliefs about the bush conflict with what we remember believing. The sudden deletion of the original interpretation is the source of the humorous effect.

Suslov has drawn on the views of Spencer, which had it that actual physical laughter was caused by the discharge of nervous energy from mental process into the muscles. This theory, once popular, suffered the defect that there really is no such thing as 'nervous energy' which behaves in this pseudo-hydraulic style; but Suslov thinks it can be at least partially resurrected if we think of the process as excess energy arising from the clearance of large sections of a neural network (when a scenario is deleted). He recognises that this is still not really an accurate biological description of the way neurons work, but he evidently still thinks there's an underlying truth in it.

One further point is necessary to the plausibility of the theory, namely that humour can be driven out by other factors. We may laugh when we realise the 'bear' is really a bush, but not when we make the reverse discovery. This is because the 'nervous energy', if we can continue to use that term, is directed into other emotions, and hence goes on to power shaking with fear rather than laughter. Suslov goes on to explain a number of other features of humour in terms of his theory with a fair degree of success.

An interesting consequence if all this were true, it seems to me, is that a network-based simulation of human consciousness would also necessarily be subject to sudden discharges. It seems to me this could go two ways. Either the successful engineers are going to notice this curious and possibly damaging property of their networks, or at some stage they are going to encounter problems (the frame problem?) which can in the end only be solved by building in a special rapid-delete facility with a special provision for the tolerance of inconsistency. Use of this facility would amount to the machine laughing.

Would it, though? There would be no need, from an engineering point of view, to build in any sound effects or 'heave with laughter' motors. Would the machine enjoy laughing, and seek out reasons to laugh? There seems no reason to think so, and it is a definite weakness of the theory that it doesn't really explain why humour is anything other than a neutral-to-unpleasant kind of involuntary shudder. Suslov more or less dismisses the pleasurable element in humour: it's more or less a matter of chance, he suggests, just as sneezing happens to be pleasant without that being the point of it. It's true that humans are good at taking pleasure in things that don't seem fun at first sight; making the capsaicin which is designed to deter animals from eating peppers into the very thing that makes them taste good, for example. But it's hard to accept that funny things are only pleasant by chance; it seems an essential feature of humour is being left on one side.

It's also possible to doubt whether all humour is a matter of conflicting interpretations. It's true that jokes typically work by suddenly presenting a reinterpretation of what has gone before. Suslov claims that tickling works in a similar way - our expectations about where the sensation is coming from next are constantly falsified. Are we also prepared to say that the sight of someone slipping on a banana skin is funny because it upsets our expectations? That might be part of it: but if conflicting interpretations are the essence of humour, optically ambiguous figures like the Necker cube should be amusing and binocular rivalry ought to be hilarious.¶There are of course plenty of technical issues too, apart from the inherent doubtfulness of whether the metaphor of 'nervous energy' can really be given a definite neurological meaning.

One aspect of Suslov's ideas ought to be testable. It's a requirement of the theory that the discarded interpretation is deleted, otherwise there is no surplus 'nervous energy'. But why shouldn't it simply recede to the status of alternative hypothesis? That seems a more natural outcome. If that were what happened, we should be ready to change our minds back as quickly as we changed them the first time: if Suslov is right and the discarded reading is actually deleted, we should find it difficult to switch back to the 'bear' hypothesis once we've displaced it with the 'bush' reading. That ought to show up in a greater amount of time needed for the second change of mind. I doubt whether experiments would find that this extra delay actually occurs.

Call Me 'Four Brains'

9 December 2007

The New Scientist had a rather rambling piece on recent ideas about the subconscious the other day. One thing I thought was interesting was the four-part model of the mind it described, attributed to Dayan, Daw and Niv. I haven't been able to find a paper which actually sets out the four-part system: the one referred to by the New Scientist is really about how control is passed between two of the four systems - but the gist is fairly clear. One of the appealing things about this line of thought is that each system has a fairly robust basis in neurology, with evidence of functions being localised in particular brain areas; but also a rationale in terms of what the system does for us and why this particular combination of systems might work well.

To begin with, we have the 'Pavlovian' system. Pavlov, of course, is famous for his work on conditioned reflexes: training dogs to salivate at the sond of a bell, and so on. Here the term is used rather loosely to cover instincts and a range of 'automatic' behaviour. It may in fact seem doubtful whether all such behaviour stems from a single system, and indeed in this case a single clear neurological location doesn't exist, though various places have a role. The chief advantage of behaviour this hard-wired is clearly speed, and the main limitation is the stereotyped nature of the behaviour: well-suited to cases where the required action is clear and urgent, and not so good elsewhere.

Then we have the habitual controller, which I think we could call the 'autopilot': responsible for more complex learned behaviour. This system can take control of behaviour for an extended period, and respond appropriately to inputs so long as they fall within the expected range; so that when driving, for example, we may take account of the progress of the car in front without having to think about it consciously. This second system can cope with much more complex problems than the Pavlovian one, but it is still largely stereotyped and when something unexpected comes up it hands over abruptly to another system.

That might be the episodic controller, which produces behaviour which makes decisions in a conscious but unreflective manner, drawing on memory of what happened before. It works well in circumstances where events are as it were, following a script and the broad outlines of what is going on are known. Although it takes a little more time and thought than either of the preceding systems, it is still fast and it has the advantage of working relatively well when detailed knowledge is not available: so long as the circumstances are broadly familiar, we can go on producing generic behaviour which is likely to be appropriate.

The fourth system is the goal-directed controller; here for the first time the brain attempts to look forward and decide which behaviour is going to achieve its goals. It is slower and consumes more resources than the other systems, and it can only really work where enough information is available, but these disadvantages are outweighed by its sophistication and flexibility. Dayan et al describe it as searching through a tree of possibilities, which I think may under-rate or misdescribe it - the brain seems to be able to devise goal-oriented strategies by some other means than working through alternatives - but that doesn't invalidate the overall model.

So if the model has a good basis in both neurology and function, how does it fare against introspection - does the mind really feel like a four-part operation from the inside? I think in many respects the model is recognisable, though I'd quibble over some details. The main weakness, as I see it, is that the perspective which the whole thing is based on is very much one in which the function of the mind is to produce good output behaviour from a range of different environmental inputs. That's fine as far as it goes, but consciousness is generally taken to be about awareness and experience as much as decision-making. When I'm sitting still and admiring the view across the bay, which system is working? None, it would appear, but in old-fashioned two-part terminology, I should have said that both conscious and subconscious were hard at work. Perhaps all the systems are working in harmony to produce the appropriate output behaviour of not doing anything?

Reincarnation

2 January 2008

Jonathan Edelmann and William Bernet have made a sterling effort to bring reincarnation within the pale of scientific investigation. Their paper, in the latest JCS, is not directly concerned with the reality of reincarnation, but with the methods which could be adopted to ensure the academic credibility of future research.

They put to one side cases where the recollection of a previous life is induced by hypnosis, because of the obvious difficulties in ensuring that the subject is not led or influenced by the hypnotist and concentrate instead on 'spontaneous' cases - those where a child comes up with memories of a former life without any particular prompting.

I say 'without any particular prompting', but the authors recognise that many alleged cases of reincarnation occur within cultures where a belief in the phenomenon is a religious obligation, and where family members are likely to prompt and encourage any signs of recollection displayed by a child. Indeed, one of their main concerns is to establish interview procedures which will eliminate direct family influence, provide an objective assessment, and include a comparison of the household described in the child's recollections with a control household.

I salute their aspiration to scientific rigour, but their efforts are tragically misplaced. Scientific investigation is wasted if the hypothesis under investigation is incoherent, and I think the notion of reincarnation is pretty much unsalvageable. It rests on a confused conception of identity, and once your ideas about identity are clarified - in any rational way - it becomes absurd. To put the problem at its most general: proponents of reincarnation accept any resemblance between dead and live persons as evidence for reincarnation- it can be personality, memories, tastes, abilities, or even physical characteristics. But if all properties are equally signs of identity, the missing resemblances are as salient as the present ones. If a gift for juggling is claimed as a sign of identity between dead A and live B in one case, I'm entitled to point out that dead X and live Y differ in their juggling abilities, though allegedly Y remembers X's life. But this is never allowed; only the points of resemblance are ever considered. Frankly, it's superstition.

I don't think that's right. The best evidence for reincarnation isn't from resemblances of that kind, but the recollection of factual information about previous lives. The ability to produce detailed memories of a former existence is so remarkable that even a few instances constitute striking evidence. The fact that some other details are not recalled does not cancel that evidence out. If I could describe my car and tell you its registration number, that would be good evidence that I really had at least seen it before: the fact that I couldn't tell you what was inside it or what brand name was on the battery wouldn't disprove that.

Edelmann and Bernet do seem to countenance a range of different evidence in principle, though: they begin by quoting the four-point SOC (Strength of Case) scale developed at Virginia University. The four points are, briefly:¶

- birthmarks/defects that correspond to the previous life;
- strength of statements about the previous life;
- relevant behaviours that relate to the previous life; and
- possible connections between present and previous life


See what I mean? Look at that first point. Birthmarks and defects? Look, reincarnation is supposed to be the transfer of a soul into another body, not the transfer of a body into a body. If birthmarks and defects are transferred, why not the disease that killed the original person? Why not the signs of old age? If physical traits are transferred, why don't babies get reincarnated as old people? In fact, why don't they come back as corpses?


You're the last person who should be surprised that mental traits can have a physical expression. I'm not asserting that birthmarks are signs of reincarnation, but if you believe in souls, there's nothing contradictory in supposing that certain physical characteristics impress themselves on the spirit in a way that others don't. and that these impressed characteristics can then be echoed in the body when the spirit arrives in a new corporeal host.

Anyway, let me finish the exposition before you start arguing - as I explained, we're addressing methodological issues here rather than the reality of reincarnation in itself.

Edelmann and Bernet propose four phases of research. In the first, the child is questioned about its earlier life in a videotaped interview conducted by professionals, who seek to draw out clear, specific and verifiable information. A second group of researchers evaluates the data gathered in phase 1, checking on the child's life and circumstances to eliminate the possibility of their having acquired information about a previous life by normal means. In this respect, the authors note that the best subjects are likely to be young children, since they are least likely to have been able to research earlier lives or pick up data by normal means of communication. The second group of researchers go on to draw up a list of 20 'descriptors' - items about the supposed previous life drawn from the interviews with the child. They also identify the site of the earlier life and another superficially like it. They might, for example, find the house where the child claims to have lived, and then pick a house with the same number in a different street.

On to phase 3. A further group of researchers is now given the two addresses (without being told which is which) and the list of descriptors: they then score both sites according to how many of the descriptors apply. In the final phase, the whole exercise is re-examined for flaws or mistakes, and the results evaluated statistically.

Sadly, no research along these lines has taken place, but it seems to hold out the possibility of opening up reincarnation for proper scientific research.


The trouble is, the research is still going to be tainted, isn't it? We're dealing with cultures where reincarnation is accepted; how can they ever eliminate the possibility that Mum

and Dad have hit on the idea that Junior is the reincarnation of Great-Uncle Jasper, and told him all sorts of stuff they know about the old man and his life? It seems hopeless to me. Maybe it would work if they could show that Junior was able to remember Uncle Jasper's bank account password, or something else that no-one else would have known.

But it's hopeless on a deeper level. What are the criteria for personal identity? Nobody really has an uncontroversial answer, so how can we begin making claims that a particular live person is identical with a certain dead one?

After all, I am not exactly the same as I was a year ago, let alone as I was when I was five: some people would say that my claim to be identical with that five year old is really only a kind of polite or convenient fiction. I wouldn't go that far: it seems to me that biological continuity is a pretty good indicator of identity. I'd be inclined to make death and birth the ending and beginning of new individuals more or less by definition - so even if you are just like Uncle Jasper, and even share some of his memories, that doesn't mean you are him.


I accept that there are issues about identity. It might be that reincarnation doesn't turn out to be what we think it is. Edelmann and Bernet suggest that if reincarnation really happens, we must transform our view of the ontology of consciousness, rule out reductive materialism, and look again at non-physical views. I think they're only partly right: it seems perfectly feasible to me to come up with a version of reincarnation which is compatible with materialism. Just to take an easy example: suppose consciousness really is a kind of electromagnetic buzz, as people like Pockett and McFadden have supposed. It seems prima facie possible that this buzz could get echoed or stored in some way and have an effect on the emergent buzz of an infant, transferring memories and personality traits - perhaps even some physical ones.

The way I see it, the first step is to establish the reality or otherwise of the transfer of memory - metemmnemonism? - along the lines Edelmann and Bernet have suggested. Then we can have the discussion about whether 'metemmnemonism' implies metempsychosis.


Gotta love your idea of rigorous scientific materialism there - no souls, just 'buzzes'.

I'd be happy with the idea of this research if I didn't know how it would go. Suppose the research takes place and finds no reliable evidence of reincarnation. Will the researchers then conclude the thing is disproved, case closed? No: they'll say they failed to find evidence, but someone should have another go. Negative results will not be counted, but when some idiot messes up the procedures and gets an invalid positive result, it'll be acclaimed and enter the mythology as cast-iron proof. That's paranormal research for you - we may not get reincarnated, but the discredited theories always come back from the dead.

What's wrong with computation?

13 January 2008

You may have seen that Edge, following its annual custom, posed an evocative question to a selection of intellectuals to mark the new year. This time the question was 'What have you changed your mind about? Why?'. This attracted a number of responses setting out revised attitudes to artificial consciousness (not all of them revised in the same direction). Roger Schank, notably, now thinks that 'machines as smart as we are' are much further off than he once believed, though he thinks Specialised Intelligences - machines with a narrow area of high competence but not much in the way of generalised human-style thinking skill - are probably achievable in the shorter term.

One remark he makes which I found thought-provoking is about expert systems, the approach which enjoyed such a vogue for a while, but ultimately did not deliver as expected. The idea was to elicit the rules which expert humans applied to a particular topic, embody them in a suitable algorithm, and arrive at a machine which understood the subject in question as well as the human expert, but didn't need years of training and never made mistakes (and incidentally, didn't need a salary and time off, as those who tried to exploit the idea for real-world business applications noted). Schank contrasts expert systems with human beings: the more rules the system learned, he says, the longer it typically took to reach a decision; but the more humans learn about a topic, the quicker they get.

Now there are various ways we could explain this difference, but I think Schank is right to see it as a symptom of an underlying issue, namely that human experts don't really think about problems by applying a set of rules (except when they do, self-consciously): they do something else which we haven't quite got to the bottom of. This is obviously a problem - as Schank says, how can we imitate what humans are doing when humans don't know what they are doing when they do it?

Another Edge respondent expressing a more cautious view about the road to true AI is none other than Rodney Brooks. Noting that the preferred metaphor for the brain has always tended to be the most advanced technology of the day - steam engines, telephone exchanges, and now inevitably computers - he expresses doubt about whether computation is everything we have been tempted to believe it might be. Perhaps it isn't, after all, the ultimate metaphor.

It seems to me that in different ways Schank and Brooks have identified the same underlying problem. There's some important element of the way the brain works that just doesn't seem to be computational. But why the hell not? Roger Penrose and others have presented arguments for the non-computability of consciousness, but the problem I always have in this connection is

getting an intuitive grasp of what exactly the obstacle to programming consciousness could be. We know, of course, that there are plenty of non-computable problems, but somehow that doesn't seem to touch our sense that we can ultimately program a computer to do practically anything we like: they're not universal machines for nothing.

One of John Searle's popular debating points is that you don't get wet from a computer simulation of rain. Actually, I'm not sure how far that's true: if the computer simulation of a tropical shower is controlling the sprinkler system in a greenhouse at Kew Gardens, you might need your umbrella after all. Many of the things AI tries to model, moreover, are not big physical events like rain, but things that can well be accomplished by text or robot output. However, there's something in the idea that computer programs simulate rather than instantiate mental processes. Intuitively, I think this is because the patterns of causality are necessarily different when a program is involved: I've never succeeded in reducing this idea to a rigorous position, but the gist is that in a computer the 'mental states' being modelled don't really cause each other directly; they're simply the puppets of a script which is really operating elsewhere.

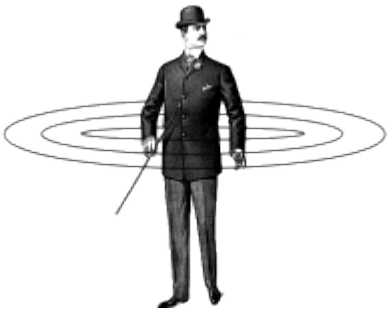
Why should that matter? You could argue that human mental states operate according to a program which is simply implicit in the structure of the brain, rather than being kept separately in some neural register somewhere; but even if we accept that there is a difference, why is it a difference that makes a difference?

I can't say for sure, but I suspect it has to do with intentionality, meaningfulness. Meaning is one of those things computers can't really handle, which is why computer translations remain rather poor: to translate properly you have to understand what the text means, not just apply a look-up table of vocabulary. It could be that in order to mean something, your mental states have to be part of an appropriate pattern of causality, which operating according to a script or program will automatically mess up. I would guess further that it has to do with a primitive form of indexicality or pointing which lies at the foundation of intentionality: if your actions aren't in a direct causal line with your behaviour, you don't really intend them, and if your perceptions aren't in a direct causal line with sensory experience, you don't really feel them. At the moment, I don't think anyone quite knows what the answer is.

If that general line of thought is correct, of course, it would be the case that we cannot ever program or build a conscious entity - but we should be able to create circumstances in which consciousness arises or evolves. This would be a blow for Asimovians, since there would be no way of building laws into the minds of future robots: they would be as free and gratuitous as we are, and equally prone to corruption and crime. On the other hand, they would also share our capacity for reform and improvement; our ability, sometimes, to transcend ourselves and turn out better than any programmer could have foreseen - to start something good and unexpected.

Egocentric Space

22 January 2008



The hypothesis that human beings have not one, but two distinct visual systems, now seems to be widely accepted. There are certainly convincing stories to be told about it both in neurological and functional terms. Neurologically, it seems that two different streams draw on the information provided by the primary visual area of the brain. A ventral stream goes eventually to the inferior temporal cortex, while a dorsal stream goes to the posterior parietal lobe. The brain being what it is, things are not quite that simple, of course, and the two streams are not completely isolated, but there seems to be good enough reason to think of them as essentially independent. Functionally, there is evidence that the brain deals separately with conscious visual perception on the one hand and 'automatic' control of actions on the other. The former system, for example, is easier to fool than the latter. When presented with certain kinds of geometrical optical illusion, we may be deceived consciously about the apparent size of an object, but when we reach out to take the object, our fingers open to just the right size anyway. There's also evidence from the effects of injuries: damage to the relevant regions of the brain may damage our conscious awareness while leaving us apparently able to use our eyes for practical purposes. Perhaps the most extreme cases are the famous instances of blindsight; subjects who cannot see at all so far as their conscious minds are concerned, but who can point to an object, or reach for one, with much greater accuracy than chance guesses could provide. Ramachandran, for one, has suggested that the two visual streams provide a satisfying answer to the puzzle of blindsight.

It is therefore an attractive hypothesis that the two streams serve regions of the brain with complementary roles, one delivering conscious visual experience and the other feeding into accurate motor control; one stream telling you what, and the other where, as some describe it. It seems to me, incidentally, that there is an interesting side-implication here about our fellow primates. Since they have the same two-stream system (in fact, I believe it was originally discovered in macaques), the implication is that they must have the same experience of doing some things deliberately, and others unthinkingly. We might have been tempted to think that animals, relying on instinct, operate on a kind of permanent autopilot, always in the same sort of state we are in when walking along without thinking of where we're going; but that seems to be ruled out at least as far as primates are concerned.

Be that as it may, a further step has been taken by many researchers, who propose that the two visual systems must encode information differently. The system which delivers conscious perception, drawing on the ventral stream, must surely encode the positions of objects allocentrically, ie in relation to the scene before us, rather than egocentrically, ie plotting distance and direction from the observer. Since this system is concerned above all with identifying things, it's surely most helpful to have things encoded in a way that doesn't change every time we move around the room, they argue. In the case of the other system, where the top priority is such tasks as deciding how far to extend your arm in order to grab something, an egocentric scheme of coding makes more sense. This view is buttressed by an argument that the allocentric coding of the perception system is what makes it more vulnerable to optical illusions.

However, in a paper for Philosophy and Phenomenological Research, Robert Briscoe now rides to the rescue of egocentricity. He thinks that our conscious perception can perhaps be regarded in this respect as an extension of proprioception, the special sense which tells us where the different parts of our body are; and both, he claims, are plausibly based on an egocentric coding scheme.

He dissociates himself from two positions which appear to have been severely undercut by the two-systems hypothesis. The first, Experience-Based Control (EBC) asserts that our fine motor control draws on the richness of our conscious perception (not all of which necessarily gets our full attention); the other, the Grounding Hypothesis, is even stronger, claiming that it is necessarily grounded in conscious experience. All such views are rejected by Briscoe, which leaves him free to sidestep the evidence which tells against them. Turning to the arguments based on optical illusion, he argues that the undeniably greater vulnerability of conscious experience might not stem from allocentric coding, but from the diversity of sources drawn on by conscious experience and hence be attributable to difficulties with integration.

And after all, he concludes, moving from defence to attack, the two systems do in fact communicate and co-operate, with conscious experience identifying the targets and the other stream sorting out the details of the movements required. If they use different coding schemes to represent the positions of objects, there are going to be some problems of translation which don't arise if both systems are egocentric.

This all seems fairly convincing to me, but I wonder whether the dispute will eventually turn out to have been misconceived: one of those either/or disputes where neither hypothesis actually catches the truth properly. Perhaps allocentric and egocentric coding aren't really the only alternatives and one or both systems operate on some mixed, intermediate, or entirely different principle. In fact, Briscoe's own views hint at this to some degree. In his view, conscious perception is an extension of proprioception; but proprioception does not, he says relate everything to some arbitrary central point - rather, it relates different body parts to each other. If we extend that approach to the external world, we seem to get a system which relates objects to each other rather than a central observer; that sounds as if it has a tinge of allocentrism about it. I accept that a clear distinction can be drawn in practice between the ability to make allocentric and egocentric judgements, but that difference doesn't have to be reflected all the way down to the level of the coding schemes involved.

Briscoe goes on to offer a few concluding remarks which perhaps reveal something of his underlying motivation: in essence, he wants to preserve a strong connection between conscious perception and agency. I sympathise with this, even with the strong claim that we need at least a potential for agency before we can have perception. But I remain to be convinced that egocentric coding of conscious visual experience is indispensable to the cause.

Phantom Penises

28 January 2008

It often happens that when someone has had a limb amputated, they experience feelings in the limb they haven't got any more - the 'phantom limb' phenomenon. The phantoms may be just temporary, a curious by-product of the operation. Sometimes the feelings are partial - where an arm has been removed, for example, patients may feel as though they still had their hand but attached to the shoulder without any intervening arm. Sometimes the experience is more of a problem, with feelings of intense pain in the amputated part which won't go away and can't be treated by normal means.

I therefore winced slightly on learning that as well as phantom limbs, there are phantom penises, experienced by those who have undergone a penectomy (a word which is well worthy of a wince in itself). V.S.Ramachandran, who devised an ingenious way of using mirrors to help people with phantom limb pain, by fooling the brain into briefly believing that the missing limb was back, has now turned his attention to penises, together with P.D. McGeoch. This time the research is not about pain relief, however, but gender identity. Penectomies, it seems, are performed for two main reasons; to eliminate a malignant cancer, or as part of gender reassignment treatment. Since male-to-female transsexuals typically feel themselves to be 'a woman in a man's body', Ramachandran and McGeoch reasoned that their response to penectomy might well be different from that of other patients. And so it proved: while 58% of men who have undergone penectomy for other reasons reported sensation in a phantom penis afterwards, only 30% of those who had done so as part of gender reassignment had a similar experience. So people who felt that a penis was not part of their true body image were much less likely to experience a phantom penis after removal.

Stranger still, perhaps, 62% of a group of female-to-male transsexuals reported having had phantom penis sensations before any surgery. In many cases the sensations dated back for years: in others, they did not occur until hormone treatment had begun. No non-transsexual women, unsurprisingly, reported the sensation of having a phantom penis ('even when prompted' as the researchers say). Ramachandran and McGeoch conclude that the study backs the view that gender identity feelings are hard-wired into the brain, and in transsexuals may be at odds with actual physical shape. They recognise, however, that there are some potential criticisms of the way the research was done.

One weak point is the risk of confabulation. By asking female-to-male transsexuals whether they had ever had phantom penis sensations, were the researchers discovering a phenomenon, or creating one? Transsexuals often have to struggle for the acceptance of their view of themselves; they have a natural reason to want to assert anything that might strengthen their case. The experience of a phantom penis would clearly be a useful piece of evidence in this context. Since female-to-male transsexuals by definition feel that they ought to have a penis, it may not be much of a leap to say they feel as though they have one, once the possibility is suggested.

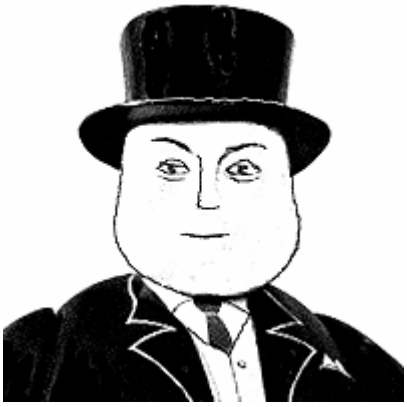
To some extent, moreover, male-to-female transsexuals might have been inclined to feel that any report of phantom sensations was letting the side down in some subtle way; suggesting that they or their bodies somehow couldn't give up the idea of having a penis very readily. Indeed, an old Freudian theory which the researchers pour scorn on, had it that the symptoms of phantom limbs expressed an unconscious desire that the limb was still there, so reasoning along those lines is by no means impossible.

However, the researchers have a number of counterarguments. Perhaps the most striking is that female-to-male transsexuals were often able to report details of the phantom penis and its behaviour, saying that it fell short of their ideal penis, for example. Surely an imagined penis, a wish-fulfilment penis, would be fully satisfactory? Less convincingly, I think, the researchers quote cases where the subjects reported the phantom penis behaving in ways - morning or unprompted erections, for example - which a female subject allegedly would have been unlikely to add to a confabulated account. I suspect the female subjects are likely to have been more aware of this kind of detail than the researchers suppose.

At the end of the day, we seem to have some suggestive evidence, but not a fully convincing case. Ramachandran and McGeoch rightly say that evidence from brain imaging studies would be very useful - notably in establishing whether a pre-operative female-to-male transsexual having a phantom penis experience has similar brain activity to a male having normal penis sensations.

Controllers In the Brain

16 February 2008



Asim Roy insists in *Connectionism, controllers and a brain theory* that there are controllers in the brain. This is not as sinister as it might sound.

Roy presents his views as an attack on connectionist orthodoxy. Connectionists, he says, believe that the brain does not have in it groups of cells that control other parts of the brain. He cites many sources, and quotes explicitly from Rumelhart, Hinton, and McClelland, who say

"There is one final aspect of our models which is vaguely derived from our understanding of brain functioning. This is the notion that there is no central executive overseeing the general flow of processing…"

Roy begins by addressing a startlingly radical argument against the notion of controllers in any system. He takes up the analogy of a human being driving a car. The steering and other systems, according to a normal understanding, are controlled by the driver; but, he says, there is an argument that this is the wrong way of looking at it. It's true that the steering is guided by input from the driver, but there is also feedback to the driver from the various systems of the car, which also determine what the driver does. The current position of the wheel and the response of the car determine how much force the driver applies to the wheel. So really the car is controlling the driver as well as vice versa, and the very idea of a controller within a system is a misapprehension.

This is obviously a slightly silly argument, but it serves to show that we need to define the notion of a controller properly. Roy suggests that the essence of a controller is that it is not dependent on inputs. The steering moves only in response to my inputs, whereas if I choose (and am tired of life, presumably) I can ignore any feedback from the car and turn the wheel any arbitrary number of degrees I settle on. Similarly, although my channel-hopping is normally guided by what appears on the TV screen, if I wish I can choose to change channels arbitrarily, or tap out a rhythm on the remote control which has nothing to do with the TV. A controller, in short can operate in different modes while a subservient system cannot. Armed with this definition, Roy argues that a connectionist network which learns by back-propagation requires an external agent to set the parameters, whether it be a human operator or another module within the overall system. I suppose it could be retorted that this controller is indeed external, and exercises its influence only during learning, but Roy would probably say that if we're modelling the human brain, these controllers would have to be taken to be part of the neural set-up. In any case, he goes on to give examples from neuroscience to show that some parts of the brain do indeed seem to operate as controllers of some other parts. I must say that this claim seems so evidently true to me that argument in its favour seems almost redundant.

Roy's conclusion is that his reasoning opens the way to a better approach, where instead of being left to local learning, the methods and weightings to be used can be dictated by a central system, at least on some occasions.

I think Roy's conclusion that there must be controllers within the brain is hard to disagree with: the question is more whether he's demolishing a straw man. What did Rumelhart, Hinton, and

McClelland actually mean? I suspect they meant to deny the existence of a central ‘homunculus’. a little man in the brain who does all the real work; and also to deny that the brain has a CPU, a place where all the data and instructions get matched together and processed. I don’t really think they meant to deny that any part of the brain ever controls any other part. I’m not sure that connectionists have ever reached the point of proposing an overall architecture for the brain, or even that that would be within the scope of the theory; rather, they just want to investigate a way of working which may be characteristic of parts of the brain.

I can imagine two ways connectionists might respond to Roy’s claims without directly contesting them. One would be to accept that the control function he describes exists but claim that it doesn’t reside in one fixed place. Different parts of the overall network might be in control at different times; it’s not that bits of the brain don’t control other parts, merely that control is sort of smeared around. But equally I can imagine a connectionist simply saying; of course we never meant to deny that that sort of control relation exists, so thanks for the clarification…

I think Roy’s attempt to define what a controller does is interesting, however. A controller, on his view, can follow the inputs, or operate without them. But surely other systems can operate without inputs, too? If I black out and cease to provide the car with control inputs, it doesn’t cease to function. It may function disastrously, but that’s also true if I exercise my controller’s right to ignore inputs and take a kind of existentialist approach to steering. You could say that in cases like the black-out one the controlled system is continuing to receive inputs - they’re just consistently zero. The point really is not operating without inputs but being able to ignore the ones you’re getting. I think what we’re grasping for here is that the inputs to a controller don’t determine the outputs. That’s interesting because it sounds like one version of free will. When I act freely, my actions were not determined by the environmental inputs. But it’s notoriously hard to explain how that could be so in a deterministic world.

If you’re a softy compatibilist about free will, you may be inclined to argue that it is to some extent a matter of degree. Where input A always gets output B, no question of freedom or controllerhood arises. But if there is a complex internal algorithm working away, such that input A may get any letter of the alphabet on different occasions, things start to look different. And if the outputs begin to show a certain kind of coherence or meaningful salience - if they begin to spell intelligible words - we might be inclined to say that the system is in some sense in control. If when we turn the wheel, the car does not respond, we might gasp metaphorically “The damn thing’s got a will of its own!”; further, if it actually directs itself down a side road and into the filling station we might seriously and literally begin to credit the car with intelligence. So far so Dennettian.

If that line of reasoning is right, controllerhood is indeed a tricky business: it might be that the only way to know whether some group of cells is a controller would be to watch it and see whether it did controllery things. And if that’s the case, maybe smeary connectionism has something in it after all.

Baby Pains

24 February 2008

 If ever you suspected that the 'hard problem' of consciousness was a recondite philosophical matter, of no significance in the real world, a recent piece in the NYT (discussed briefly by our old friend Steve Esser) should convince you otherwise.

It explains how Kanwaljeet Anand discovered that new-born babies were receiving surgery without anaesthetics, and goes on to discuss the evidence that fetal surgery, increasingly common, also causes pain responses in the unborn child.

Why would doctors be so callous? They don't believe new-born children, let alone fetuses, feel any pain. When they operate on a new-born baby, there may be some reflexes, but there are no pain qualia - no pain is really being felt. So all they need to give in the way of drugs is a paralytic - something that makes the subject keep still during the operation, but has no effect on how it feels.

It seems to me that, although they may not be clearly aware of it, the doctors here are making important real world decisions on the basis of their philosophical beliefs about qualia. Quite bold beliefs too: they don't believe the fetuses can be feeling pain because the brain, and in particular the cortex, is not yet wired up sufficiently for consciousness, and if you're not conscious, you can't have qualia.

It's quite possible they are right, I suppose, but I think most people who have been through some of the philosophical arguments would doubt whether we can speak of these matters with such certainty, and would be inclined to err on the side of safety. To be honest, it's a bit hard to shake the suspicion that the real reason fetuses don't get anaesthetics is because fetuses don't complain...

 I think the real issue here is slightly different. Why was Anand concerned about the babies who were coming back to him after surgery, before he even knew they weren't being anaesthetised? Because they were traumatised. Their systems had been flooded with hormones usually associated with pain, their breathing was poor, their blood sugar was out of order. When they were given anaesthetics, instead of just paralytics, they survived the operations in better health. There were medical reasons for avoiding the pain, not just subjective ones. The problem arose out of the fact that the surgeons and anaesthetists had got used to the idea that relief of subjective pain was all that mattered. So when they were dealing with patients who they judged incapable of subjective pain, they saw no reason to anaesthetise. They were, in fact, paying too much attention to the idea of qualia, and not enough to the physical, medical events which actually constitute pain even if your cortex isn't working. (The passage in the NYT piece about people with hydranencephaly is a dreadful red herring, by the way: it just isn't true that they have no cortex.)

 True, there were other reasons for using anaesthetics in that particular case. But where do the doctors get their certainty that no pain is being felt? You can't talk about your qualia if your cortex isn't working, but since qualia are an utter mystery and separable in principle from the physical operation of the brain altogether, there's no strong reason to think you're not feeling them. We take winces and grimaces as good indicators of pain in normal people: how can we be sure it's any different in foetuses? Is it just the talking that matters? Is an unreportable pain no pain at all? Surely not.

 You say qualia are separable from the physical operation of the brain, but you don't really believe that in any important way. You know quite well that brain events like the firing of certain neurons are what cause these sensations you misdescribe as ineffable qualia. If I hit you with a physical spade, you'll feel pain qualia alright, however 'separable' you think they are.

Let's get a bit more philosophical here. Why is pain bad? Why is it to be avoided and minimised? There are several traditional answers -- I'd say the following just about cover it. ¶¶

- It just is. Pain is just bad in its essential nature.
- The moral rules agreed by our society enjoin us not to cause pain.
- Ethics requires us to seek the greatest possible balance of pleasure over pain.
- You wouldn't want pain inflicted on you, so don't inflict it on other people.
- Witnessing or hearing about pain makes me feel bad.
- Carelessness about pain is the beginning of a general callousness which might undermine our concern for others, without which we're doomed.

The first doesn't mean anything, if you ask me. The second might be right, but the established rules, judging by medical practice, seem to say foetal pain is something we don't have to worry about. Social rules are negotiable, anyway, so there's no final answer there. The third one, like the first one, just assumes pain is bad. The fourth is OK but begs the question so far as foetuses are concerned, because we don't know what I'd want done to me if I were a foetus. I might just want them to get on with the operation, unbothered by the apparent 'pain' I wasn't actually feeling. Numbers five and six, if you ask me, get close to the real motivation here.

Got that? OK, now let's revert to your question -- why do the doctors behave like this, why do they torture babies? It's not just, I suggest, that they don't believe in foetal pain. The more crucial point is that they know foetuses don't remember pain. There was a time, in the not-too-distant past, when some children, not just babies, were merely given curare before being operated on. Just like the new-born babies here, it paralysed them so the surgeon could get on with the job but did nothing whatever to relieve the pain. Did the doctors think the children didn't feel the pain? Well, there's some talk about them thinking curare was anaesthetic, but that's balderdash -- they didn't use it for adults. No, I believe they knew quite well the children were in pain, they just thought it didn't matter because children don't remember these things. Give 'em an ice cream, they thought, and we'll hear no more about it. (I'm not saying that's necessarily correct, by the way).

In essence, the doctors implicitly agreed with me. There were no deep metaphysical or ethical reasons to avoid pain, and reciprocity never really worked, because there were always relevant differences between you and the person suffering the pain (unless you were fighting your twin brother, perhaps). You could always say; yeah, do as you would be done by, but if I were in the

state that person's in, I would want it done to me. So, the doctors rightly concluded, there were only two fundamental reasons to avoid causing pain; first, it upset people. As a result, our social rules were generally set to minimise it, and so second, the unnecessary infliction of pain would tend to have bad social effects, undermining the rules and generally risking a withdrawal of social consent. Pain which will never be remembered cannot upset us or have bad social consequences -- so it doesn't matter.

There's more evidence that this is the established medical view. Besides paralysing and anaesthetising patients, doctors use drugs which specifically remove the memory. In some cases, drugs which remove the memory have been used instead of anaesthetics, just because in medical eyes, pain you don't remember is the same as pain that never happened. It's perfectly normal contemporary practice to use a mixture of true anaesthestics and amnestics.

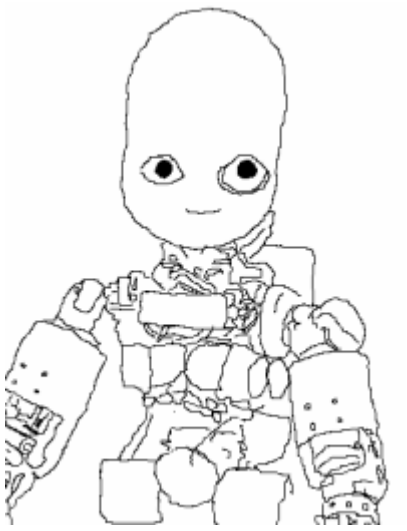
So, to sum up, pain has three aspects: medical (the hormonal reaction, the changes in the body), psychological (we don't like thinking about it), and social (we've agreed to outlaw pain, and erosion of that rule undermines society generally). And that's all there is. The medical, psychological, and social aspects add up to what pain is: there's no mystical component, no qualia.


That just seems mad to me - bonkers and almost diabolical. Surely you can see that the reason pain is bad is because it hurts? That's all pain is - hurting.

You and your supposed doctor friends are a bit over-confident about your amnesia anyway, aren't you? The NYT article reports evidence that pain experienced early on affects a child's responses later. Do you really feel confident that agonising episodes early on -- even in the womb -- are not lurking in some damaging form in the subconscious of the child, or even the adult?

iCub iTalk?

4 March 2008



More about babies - of a sort. You may have seen reports that Plymouth University, with support from many other institutions, has won the opportunity to teach a baby robot to speak. The robot in question is the iCub, and the project is part of Italk, funded by the EU under its Seventh Framework Programme, to the tune of £4.7 million (you can't help wondering whether it wouldn't have been value for money to have slipped Steve Grand half a million while they were at it...).

The gist of it seems to be that next year the people at Plymouth will get the iCub to engage in various dull activities like block-stacking (a perennial with both babies and AI) and try to introduce speech communication about the tasks. It is meant to be learning in a way far closer to the typical human experience than anything attempted before. Unfortunately, I haven't been able to find any clear statement of how they expect the language skills to work, though there is quite a lot of technical detail available about the iCub. This is evidently a pretty splendid piece of kit, although the current model has a mask for a face, which means none of the potent interactive procedures which depend on facial expression, as explored by Cynthia Brezeal and others, will be available. This is a shame, since real babies do use face recognition and smiles to get more feedback out of adults.

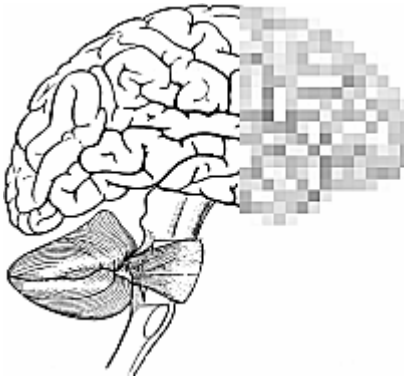
In one respect the project has an impeccable background, since Alan Turing, in the famous paper which arguably gave rise to modern Artificial Intelligence, speculated that a thinking robot might have to begin as a baby and be educated. But it seems a tremendously ambitious undertaking. If we are to believe Chomsky, the human language acquisition device is built in - it has to be, since human babies get such poor input, with few corrections and limited samples, yet learn at a fantastic speed. They just don't get enough information about the language around them to be able to reverse-engineer its rules; so they must in fact simply be customising the setting of their language machine and filling up its vocabulary stores. The structures of real-world languages arguably support this point of view, since the variations in grammar seem to fall within certain limited options, rather than exploiting the full range of logical possibilities. If this is all true, then a robot which doesn't have a built-in language facility is not going to get very far with talking just by playing with some toys.Of course Chomsky is not, as it were, the only game in town: a more recent school of thought says that by treating language as a formal code, and assuming that babies have to learn the rules before they can work out what people mean, Chomsky puts the cart before the horse; actually, it's because babies can see what people mean that they can crack the code of grammar so efficiently.

That's a more congenial point of view for the current project, I imagine, but it raises another question. On this view, babies are not using an innate language module, but they are using an innate ability to understand, to get people's drift. I don't think the Plymouth team have worked out a way of building understanding in beforehand (that would be a feat well worth trumpeting in its own right), so is it their expectation that the iCub will acquire understanding through training? Or are their aims somewhat lower?

It seems a key question to me: if they want the robot to understand what it's saying, they're aiming high and it would be good to know what the basis for their optimism is (and how they're going to demonstrate the achievement). If not, if they're merely aiming for a basic level of performance without worrying about understanding (and the selection of experiments does rather point in that direction), the project seems a bit underwhelming. Would this be any different, fundamentally, from what Terry Winograd was doing, getting on for forty years ago (albeit with SHRDLU, a less charismatic robot)?

Blue Brain – Success?

8 March 2008



A bit of an update in Seed magazine on the Blue Brain project. This is the project that set out to simulate the brain by actually reproducing it in full biological detail down to the behaviour of individual neurons and beyond: with some success, it seems.

The idea of actually simulating a real brain in full has always seemed fantastically ambitious, of course, and in fact the immediate aim was more modest: to simulate one of the columnar structures in the cortex. This is still an undertaking of mind-boggling complexity: 10,000 neurons, 30 million synaptic connections, using 30 different kinds of ion channel. In fact it seems the ion channels were one of the problem areas; in order to get good enough information about them, the project apparently had to set up its own robotic research. I hope the findings of this particular bit of the project are being published in a peer-reviewed journal somewhere.

However, the remarkable thing is that it worked: eventually the simulated column was created and proved to behave in the same way as a real one. So is the way open for a full brain simulation? Not quite. Even setting aside the many structural challenges which surely remain to be unravelled (don't they - the brain isn't simply an agglomeration of neocortical columns?) Henry Markram, the project Director, estimates that an entire brain would require the processing of 500 petabytes of data, way beyond current feasibility. Markram believes that within ten years, the inexorable increase in computing power will make this a serious possibility. Maybe: it doesn't pay to bet against Moore's Law - but I can't help noticing that there has been a big historical inflation in the estimated need, too. Markram now wants 500 petabytes: a single petabyte is 1015 bytes; but in 1950 Turing thought that 1015 bits represented the highest likely capacity of the brain, with about 109 enough for a machine which could pass the Turing Test. OK, perhaps not really a fair comparison, since Turing had nothing like Blue Brain in mind.

One criticism of the project asks how it judges its own success - or rather, suggests that the fact that it does judge its own success is a problem. If we had a full brain which could operate a humanoid robot and talk to us, there would be no doubt about the success of the project; but how do we judge whether a simulated neuronal column is actually working? The project team say that their conclusions are based on scrupulous comparisons with real biological brains, and no doubt that's right; but there's still a danger that the simulation merely confirms the expectations fed into it. They came up with an idea of how a column works; they built something that worked like that: and behold, it works how they think a column works.

There is also undeniably something a bit strange about the project. Before Blue Brain was ever thought of, proponents of AI would sometimes use the idea of a total simulation as a kind of thought-experiment to establish the merely neurological nature of the mind. OK, there might be all these mechanisms we didn't understand, and emergent phenomena, and all the rest, but at the end of the day, what if we just simulated everything? Surely then you'd have to admit, we would have made an artificial mind - and what was to stop us, except practicality? It was an unexpected development back in 2005 when someone actually set about making this last-ditch argument a reality. It is unique; I can't think of another case where someone set out to

reproduce a biological process by building a fully detailed simulation, without having any theory of how the thing worked in principle.

This raises some peculiar possibilities. We might put together the full Blue Brain; it might be demonstrably performing like a human brain, controlling a robot which walked around and discussed philosophy with us, yet we still wouldn't know how it did it. Or, worse perhaps, we might put it all together, see everything working perfectly at a neuronal level, and yet have our attached robot standing slack-jawed or rolling around in a fit, without our being able to tell why.

It may seem unfair to describe Markram and his colleagues as having no theory, but some of his remarks in the article suggest he may be one of those scientists who doesn't really get the Hard Problem at all.

...It's the transformation of those cells into experience that's so hard. Still, Markram insists that it's not impossible. The first step, he says, will be to decipher the connection between the sensations entering the robotic rat and the flickering voltages of its brain cells. Once that problem is solved—and that's just a matter of massive correlation—the supercomputer should be able to reverse the process. It should be able to take its map of the cortex and generate a movie of experience, a first-person view of reality rooted in the details of the brain...

It could be that Markram merely denies the existence of qualia, a perfectly respectable point of view; but it looks as if he hasn't really grasped what they are, and that they can't be captured on any kind of movie. Perhaps this outlook is a natural or even a necessary quality of someone running this kind of project. I suppose we'll have to wait and see what happens when he gets his 500 petabyte capacity.

Four Zomboids

23 March 2008

I've suggested previously that one of the important features of qualia (the redness of red, the smelliness of Gorgonzola, etc) is haecceity, thisness. When we experience redness, it's not any Platonic kind of redness, it's not an idea of redness, it's that. When people say there is something it is like to smell that smell, we might reply, yes: that. That's what it's like. The difficulty of even talking about qualia is notorious, I'm afraid.

But it occurred to me that there was another problem, not so often addressed, which has the same characteristic: the problem of one's own haecceity.

Both problems are ones that often occur to people of a philosophical turn of mind, even if they have no academic knowledge of the subject. People sometimes remark 'For all we know, when I see the colour blue, you might be seeing what I see when I see red', and similarly, thoughtful people are sometimes struck by puzzlement as to why they are themselves and not someone else. That particular problem has a trivial interpretation - if you weren't you, it would be someone else's problem - but there is a real and difficult issue related to the grand ultimate question of why there is anything at all, and why specifically this.

One of the standard arguments for qualia, of course, is the possibility of philosophical zombies, people who resemble us in every way except that they have no qualia. They talk and behave just like us, but there is nothing happening inside, no phenomenal experience. Qualophiles contend that the possibility of such zombies establishes qualia as real and additional to the normal physical story. Can we have personhood zombies, too? These would be people who, again, are to all appearances like us, but they don't have any experience of being anyone, no sense of being a particular self. It seems at least a prima facie possibility.

That means that if we consider both qualia and selfhood, we have a range of four possible zomboid variants. Number one, not in fact a zombie at all, would have both qualia and the experience of selfhood - probably what the majority would consider the normal state of affairs. His opposite would have neither qualia nor a special sense of self, and that would be what a Dennettian sceptic takes to be the normal position. Number three has a phenomenal awareness of his own existence, but no qualia. This is what I would take to be the standard philosophical zombie. This is not really clear, of course: I assume the absence of discussion of the self in normal qualia discussion implies that zombies are normal in this respect, but others might not agree and some might even be inclined to regard the sense of self as just a specific example of a quale (there are, presumably, proprioceptive qualia, though I don't think that's what I'm talking about here), not really worthy of independent discussion.

It's number four that really is a bit strange; he has qualia going on inside, but no him in him: phenomenal experience, but no apparent experiencer. Is this conceivable? I certainly have no knock-down argument, but my inclination is to say it isn't: I'm inclined to say that all experiences have an experiencer just as surely as causes have effects. If that's true, then it suggests the two cases of haecceity might be reducible to one: the thisness of your qualia is

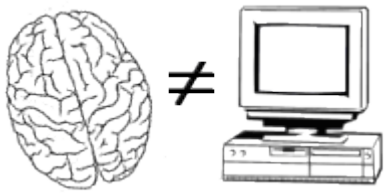
really just your own thisness as the experiencer (I hope you're still with me here, reader). That in turn might mean we haven't been looking in quite the right place for a proper account of qualia.

What if number four were conceivable? If qualia can exist in the absence of any subjective self-awareness, that suggests they're not as tightly tied to people as we might have thought. That would surely give comfort to the panpsychists, who might be happy to think of qualia blossoming everywhere, in inanimate matter as well as in brains. I don't find this an especially congenial perspective myself, but if it were true, we'd still want to look at personal thisness and how it makes the qualia of human beings different from the qualia of stones.

At this point I ought to have a blindingly original new idea of the metaphysics of the sense of self which would illuminate the whole question as never before. I'm afraid I'll have to come back to you on that one.

Not the Same

4 April 2008



Chris Chatham has a nice summary of ten key differences between brains and computers which is well worth a read. Briefly, the list is:¶¶

- Brains are analogue; computers are digital¶
- The brain uses content-addressable memory¶
- The brain is a massively parallel machine; computers are modular and seria¶
- Processing speed is not fixed in the brain; there is no system clock¶
- Short-term memory is not like RAM¶
- No hardware/software distinction can be made with respect to the brain or mind¶
- Synapses are far more complex than electrical logic gates¶
- Unlike computers, processing and memory are performed by the same components in the brain¶
- The brain is a self-organizing system¶
- Brains have bodies¶
- The brain is much, much bigger than any [current] computer¶

¶Actually eleven differences - there's a bonus one in there.

Hard to argue with most of these, and there are some points among them which are well worth making. There is always a danger, when comparing the capacities of brains and computers, of assuming a similarity even when trying to point up the contrast. There have, for example, been many attempts to estimate the storage capacity of the brain, or the memory, over the years for example, always concluding that it is huge; but a figure in megabytes doesn't really make much sense. Asking how many bytes there are in the memory is like asking how many pixels Leonardo needed to do the Mona Lisa: it's not like that. Chris generally steers just clear of this danger, although I'd be more inclined to say, for example, that the concept of processing speed has no useful application in the brain rather than that it isn't fixed.

I wonder a bit about some of the positive assertions he makes. Are brains analogue? Granted they're not digital, at least not in the straightforward way that a digital computer is, but unless we take 'analogue' as a synonym for 'non-digital' it's not really clear to me. I take digital and analogue to be two different ways of representing real-world quantities; I don't think we really know exactly how the brain represents things at the moment. It's possible that when we do know, the digital/analogue distinction may seem to be beside the point. And are brains 'massively parallel'? It's a popular phrase, but one of the older bits of this site long ago looked at a few reasons why the resemblance to massively parallel computing seems slight. In fairness, when people make this assertion, they aren't really saying that the brain is like a parallel processing set-up; they're trying to describe a quality of the brain for which there is no good word, ie that things seem to be going on all over it at the same time. Chris is really warning against an excessively modular view. Once again, we can agree that the brain is unlike computers - this time in the way they funnel data and instructions together in one or more processors; but the positive comparison is more problematic.

There's some underlying scope for confusion, too, about what we mean when we assert that brains are not computers. We could just intend the trivial point that there isn't actually a

physical PC in our heads. More plausibly, we could mean that the brain doesn't have the same general architecture and functional features as a computer (which I think is about what Chris means to do). We could mean that the brain doesn't do stuff that we could easily recognise as computation, although it might be functionally similar in the sense of producing the same output to input relationships as a computed process. We might mean it doesn't do stuff that we could easily recognise as computation, and that there appears to be no non-trivial way of deriving algorithms which would do the same thing. We might go one further and assert, as Roger Penrose does, that some of what the brain is doing is non-computable in the same sort of way as the tiling problem (though here again we have to ask is it really like that since the question of computability seems to assume the brain is typically solving problems and seeking proofs). Finally, we could be saying that the brain has altogether mysterious properties of free will and phenomenal experience which go beyond anything in our current understanding of the physical world, and ergo far beyond anything a mere computer might possess.

A good thought-provoking discussion, in any case.'

PS: Chris Chatham gamely picked up on my remarks about his interesting earlier piece, which set out a number of significant differences between brains and computers. We are, I think, somewhere between 90 and 100 percent in agreement about most of this, but Chris has come back with a defence of some of the things I questioned. So it seems only right that I should respond in turn and acknowledge some overstatement on my part.

Chris points out that "processing speed" is a well-established psychometric construct. I must concede that this is true: the term is used to cover various sound and useful measures of speed of recognition and the like: so I was too sweeping when I said that 'processing speed' had no useful application to the brain. That said, this kind of measurement of humans is really a measurement of performance rather than directly of the speed of internal working, as it would be when applied to computers. Chris also mentions some other speed constraints in the brain - things like rate of firing of neurons, speed of transmission along nerves - which are closer to what 'processing speed' means in computers (though not that close, as he was saying in the first place). In passing, I wonder how much connection there is between the two kinds of speed constraint in humans? The speed of performance of a PC is strongly affected by its clock speed; but do variations in rates of neuron firing have a big influence on human performance? It seems intuitively plausible that in older people neurons might take slightly longer to get ready to fire again, and that this might make some contribution to longer reaction times and the like (I don't know of any research on the subject), but otherwise I suspect differences in performance speed between individual humans arise from other factors.

In a nutshell, Chris is right when he says that I am probably uncomfortable with equating "sparse distributed representation" with "analog,". To me it looks like a whole 'nother thing from what used to go on in analog computers, where a particular value might be represented by a particular level of current. The whole topic of mental representation is a scary one for me in any case: if Chris wanted a twelfth difference to add to his list, I think he could add Computers don't really do representation. That may seem an odd thing to say about machines that are all about manipulating symbols; but nothing in a computer represents anything except in so far as a human or humans have deemed or designed it to represent something. Whereas the human brain somehow succeeds in representing things to itself, and then to other humans, and manages to confer representational qualities on noises, marks on paper - and computers. This surely remains one of the bogglingly strange abilities of the mind, and it's unlikely the computer analogy is going to help much in understanding it.

I do accept that in some intuitive sense the brain can be described as ‘massively parallel’ - people who so describe it are trying to put their finger on a characteristic of the brain which is real and important . But the phrase is surely drawn from massively parallel computing, which really isn’t very brain-like. ‘Parallel’ is a good way of describing how different bits of a process can be shepherded in an orderly manner through different CPUs or computers and then reintegrated; in the brain, it looks more as if the processes are going off in all directions, constantly interfering with each other, and reaching no definite conclusion. How all this results in an ordered and integrated progression of conscious experience is of course another notorious boggler, which we may solve if we can get a better grasp of how this ‘massively parallel’ - I’d rather say ‘luxuriantly non-linear’ way of operating works. My fear is that use of the phrase ‘massively parallel’ risks deluding us into thinking we’ve got the gist already.

Whatever the answer there, Chris’s insightful remarks and the links he provided have certainly improved my grasp of the gist of things in a number of areas, however, for which I’m very grateful.

Unconscious Decisions

17 April 2008

Benjamin Libet's experimental finding that decisions had in effect already been made before the conscious mind became aware of making them is both famous and controversial; now new research (published in a 'Brief Communication' in Nature Neuroscience by Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze and John-Dylan Haynes) goes beyond it. Whereas the delay between decision and awareness detected by Libet lasted 500 milliseconds, the new research seems to show that decisions can be predicted up to ten seconds before the deciders are aware of having made up their minds.

The breakthrough here, like so many recent advances, comes from the use of fMRI scanning. Libet could only measure electrical activity and had to use the Readiness Potential (RP) as an indicator that a decision had been made: the new research can go much further, mapping activity in a number of brain regions and applying statistical pattern recognition techniques to see whether any of it could be used to predict the subject's decision.

The design of the experiments varied slightly from Libet's original ingenious set-up. This time a series of letters was displayed on a screen. The subjects were asked to press either a right or a left button at a moment chosen by them; they then identified the letter which had been displayed at the moment they felt themselves deciding to press either right or left. In the main series of experiments, no time constraints were imposed.

Two regions proved to show activity which predicted the subject's choice: primary motor cortex and the Supplementary Motor Area (SMA) - the SMA is the source of the RP which Libet used in his research. In the SMA the researchers found activity which predicted the decision some five seconds before the moment of conscious awareness, but it was elsewhere that the earliest signs appeared - in the frontopolar cortex and the precuneus. Here the subject's decision could be seen as much as seven seconds ahead of time: allowing for the delay in the fMRI response, this tots up to a real figure of ten seconds. One contrast with earlier findings is that there was no activation of the dorsolateral prefrontal cortex: the researchers hypothesise that this was because the design of their experiment did not require subjects to remember previous button presses. Another difference, of course, is the huge delay of five seconds in the SMA, which one would have expected to be comparable with the findings of earlier, RP-based research. Here the suggested explanation is that in the new experiments the timing of button presses was wholly unconstrained so that there was more time for activity to build up. The time delay in the fMRI study apparently means that the possibility that there was additional activity within the last few hundred milliseconds cannot be excluded: I conjecture that this offers another possible explanation if the RP studies actually detected a late spike which the fMRI couldn't detect.

The experimenters also ran a series of experiments where the subject chose left or right at a pre-determined time: this does not seem to have shortened the delays, but it showed up a difference between the activation in the frontopolar cortex and the precuneus: briefly, it looks as if the former peaks at the earliest stage, with the precuneus 'storing' the decision through more continuous activation.

What is the significance of these new findings? The researchers suggest the results do three things: they show that the delay is not confined to areas which are closely associated with motor activity, but begins in 'higher' areas; they demonstrate clearly that the activity relates to identifiable decisions, not just general preparation; and they rule out one of the main lines of attack on Libet's findings, namely that the small delay observed is a result of mistiming, error, or misunderstanding of the chronology. That seems correct - a variety of arguments of differing degrees of subtlety have been launched against the timings of Libet's original work. Although Libet himself was scrupulous about demonstrating solid reasons for his conclusions, it always seemed that a delay of a few hundred milliseconds might perhaps be attributable to some sort of error in the book-keeping, especially since timing a decision is obviously a tricky business. A delay of ten seconds is altogether harder to explain away.

However, it seems to me that while the new results close off one line of attack, they reinforce another - the claim that these experiments do not represent normal decision making. We do not typically make random decisions at a random moment of our choosing, and it can therefore fairly be argued that the research has narrower implications than might appear, or even that they are merely a strange by-product of the peculiar mental processes the subjects were asked to undertake. While the delay was restricted to half a second, it was intuitively believable that all our normal decisions were subject to a similar time-lag - surprising, but believable. A delay of ten seconds in normal conscious thought is not credible at all; it's easy to think of cases where an unexpected contingency arises and we act on it thoughtfully and consciously within much shorter periods than that.

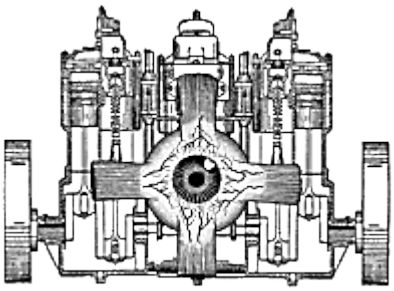
The researchers might well bite the bullet so far as that goes, accepting that their results show only that the delay can be as long as ten seconds, not that it invariably is. Libet himself, had he lived to see these results might perhaps have been tempted to elaborate his idea of 'free won't' - that while decisions build up in our brains for a period before we are aware of them, the conscious mind retains a kind of veto at the last moment.

What would be best of all, of course, is further research into decisions made in more real-life circumstances, though devising a way in which decisions can be identified and timed accurately in such circumstances is something of a challenge.

In the meantime, is this another blow to the idea of free will generally? The research will certainly hearten hard determinists, but personally I remain a compatibilist. I think making a decision and becoming aware of having made that decision are two different things, and I have no deep problem with the idea that they may occur at different times. The delay between decision and awareness does not mean the decision wasn't ours, any more than the short delay before we hear our own voice means we didn't intend what we said. Others, I know, will feel that this relegates consciousness to the status of an epiphenomenon.

Consciousness as Hardware

11 May 2008



Jan-Markus Schwindt put up a complex and curious argument against physicalism in a recent JCS: one of those discussions whose course and conclusion seem wildly wrong-headed, but which provoke interesting reflections along the way.

His first point, if I followed him correctly, was that physics basically comes down to maths. In the early days, scientists thought of themselves as doing sums about some basic stuff which was out there in the world; as the sums and the theories behind them got more sophisticated and comprehensive, there was less and less need for the stuff-in-itself; the equations were really providing the whole show. In the end, the inescapable conclusion is that the world described by physics is a mathematical structure.

Schwindt quotes several scientists along these lines, such as Eddington saying 'We have chased the solid substance from the continuous liquid to the atom, from the atom to the electron, and there we have lost it'. There's no doubt, of course that maths is indeed the language, or perhaps the soul, of physics. Hooke claimed that Newton had stolen the idea of the inverse square law of gravity from him; but Christopher Wren gently remarked that he too had thought of the idea independently; so had many others he could name; the point was that only Newton could do the maths which turned it into a verifiable theory. Thes days, of course, it's notoriously hard to describe the 'stuff' that modern quantum physics is about in anything other than mathematical terms, and one respectable point of view is that there's no point in trying.

But I don't think physicalists are necessarily committed to the view that the world is mathematical to the core. In fact mathematicians have a noticeable tendency to be Platonists, believing that numbers and mathematical entities have a real existence in a world of their own. This is a form of dualism, and hence in opposition to physicalism, but very different to Schwindt's point of view - instead of merging the physical into the mathematical, it keeps the precious formulae safe and unsullied in an eternal world of their own.

Moreover, at the risk of sounding like a bad philosopher, it all depends what you mean by mathematics. Numbers and formulae can be used in more than one way. n=1 is a testable statement in arithmetic; in many programming languages it's true by fiat - it makes the variable n equal one. In Searlian terms, the direction of fit is different, or to put it less technically, in arithmetic it's a statement, in programming an instruction. Now the kind of mathematical laws which hypothetically sustain the world would have to have the same direction of fit as the programming instruction: that is, it would be up to the world to resemble the law rather than up to the law to resemble the world. In fact they would require some more advanced metaphysical power than any mere high-level programming language; the sort of force which Adam's words are supposed to have had in the garden of Eden, where his naming of a beast determined its very essence. If we could explain how apparently arbitrary laws of physics come to have this power over reality, it would be an unprecedented advance in metaphysics and the philosophy of science.

So when Schwindt tells us that the world is a mathematical structure, he may be right if he means mathematical in the sense of embodying a set of mysteriously potent ontological

commands: but that's just as complex and difficult as a world where the laws of physics are mere descriptions and the driving force of ontology is hidden away in ineffable stuff-in-itself. If, on the other hand, he means the world is reducible to a mathematical description, I think he's mistaken - or at the least, he needs another argument to show how it could be.

For a second key point, Schwindt draws on an argument against functionalism proposed by Zuboff. This says that from a functionalist point of view, consciousness just consists of the firing of a certain pattern of neurons. But look, we could take out one neuron, put it in a dish somewhere, and then hand-feed the outputs and inputs between it and the rest of the brain in such a way that it functioned normally. Surely the relevant state of consciousness would be unaffected? If you're happy with that, then you must accept that we could do the same with all the neurons in a brain. But if that's so, we could put together a set of neurons which actually belong to different brains, but which together instantiate the right pattern for a particular experience, even though they don't belong to any physically coherent individual? But surely this is absurd - so functionalism must be false.

Schwindt accepts this argument but gives it a different spin: for there to be patterns, he suggests, including patterns of neuron firing, there has to be an interpreter: someone who sees them as patterns. Since the world as described by physics is a mathematical structure, and since no mere mathematical structure can be an interpreter, consciousness requires something over and above the normal physical account. In fact, he proposes that consciousness is non-physical, and in effect the hardware which (if I've got this right) generates experience out of the mathematical structure of the physical world. This reasoning seems to lead naturally towards dualism, but Schwindt wants to stop just short of that: there seem to be two alternative reductions of the world on offer, he says, reasonably enough: the physicalist one and another which reduces everything to direct experience; but ultimately we can hope that there is some as yet unknown level on which monism is reasserted.

Alas, I think the Zuboffian argument (I haven't read the original, so I rely on Schwindt's account) fails because credible forms of functionalism don't just require a pattern of neuron firing to exist: they require a relevant pattern of causal relations between neuron firing. As soon as hand-simulation enters the picture, all bets are off.

I think one reason Schwindt takes such an unlikely route is that he goes somewhat overboard in asserting the primacy of direct experience: he's quite right that in one sense it's all we've got; but like the idealists, he is in danger of mistaking an epistemological for an ontological primacy. It won't come as any surprise, in any case, that I don't really see how consciousness can be likened to hardware. Consciousness is content; arguably that's all it is: whereas hardware is what gets left behind when you take the content out of a brain (or a computer). Isn't it? Schwindt goes further and within consciousness has qualia playing a role analogous to a monitor, but I found that idea more confusing than illuminating.

Could consciousness be hardware? I can't reject the idea: but that's only because I don't think I can properly grasp it to begin with.

The Schemata of Ouroboros

1 June 2008

Knud Thomsen has put a draft paper describing his 'Ouroboros Model' - an architecture for cognitive agents - online. It's a resonant title at least - as you may know, Ouroboros is the 'tail-eater'; the mythical or symbolic serpent which swallows its own tail, and in alchemy and elsewhere symbolises circularity and self-reference.

We should expect some deep and esoteric revelations, then; but in fact Thomsen's model seems quite practical and unmystical. At its base are schemata, which are learnt patterns of neuron firing, but they are also evidently to be understood as embodying scripts or patterns of expectations. I take them to be somewhat similar to Roger Schank's scripts, or Marvin Minsky's frames. Thomsen gives the example of a lady in a fur coat; when such a person enters our mind, the relevant schema is triggered and suggests various other details - that the lady will have shoes (for that matter, indeed, that she will have feet). The schemata are flexible and can be combined to build up more complex structures.

In fact, although he doesn't put it quite like this, Thomsen's model assumes that each mind has in effect a single grand overall schema unrolling within it. As new schemas are triggered by sensory input, they are tested for compatibility with the others in the current grand structure through a process Thomsen calls consumption analysis. Thomsen sees this as a two-stage process - acquisition, evaluation, acquisition, evaluation. He seems to believe in an actual chronological cycle which starts and stops, but it seems to me more plausible to see the different phases as proceeding concurrently for different schemata in a multi-threaded kind of way.

Thomsen suggests this model can usefully account for a number of features of normal cognitive processes. Attention, he suggests, is directed to areas where there's a mismatch between inputs and current schemata. It's certainly true that attention can be triggered by unexpected elements in our surroundings, but this isn't a particularly striking discovery, or one that only a cyclical model can account for - and it doesn't explain voluntary direction of attention, nor how attention actually works. Thomsen also suggests that emotions might primarily be feedback from the consumption analysis process. The idea seems to be that when things are matching up nicely we get a positive buzz, and when there are problems negative emotions are triggered. This doesn't seem appealing. For one thing, positive and negative reinforcement is at best only the basis for the far more complex business of emotional reactions; but more fatally it doesn't take much reflection to realise that surprises can be pleasant and predictability tediously painful.

More plausibly, Thomsen claims his structure lends itself to certain kinds of problem solving and learning, and to the explanation of certain weaknesses in human cognition such as priming and masking, where previous inputs condition our handling of new ones. He also suggests that sleep fits his model as a time of clearing out 'leftovers' and tidying data. The snag with all these claims is that while the Ouroboros model does seem compatible with the features described, so are many other possible models; we don't seem to be left with any compelling case for adopting the serpent rather than some other pattern-matching theory. The claim that minds

have expectations and match their inputs against these expectations is not new enough to be particularly interesting: the case that they do it through a particular kind of circular process is not really made out.

What about consciousness itself? Thomsen sees it as a higher-order process - self-awareness or cognition about cognition. He suggests that higher order personality activation (HOPA) might occur when the cycle is running so well there is, as it were, a surfeit of resources; equally it might arise when when a particular concentration of resources comes together to deal with a particularly bad mismatch. In between the two, when things are running well but not flawlessly, we drift on through life semi-automatically. In itself that has some appeal - I'm a regular semi-automatic drifter myself - but as before it's hard to see why we can't have a higher-order theory of consciousness - if we want one - without invoking Thomsen's specific cyclical architecture.

In short it seems to me Thomsen has given us no great reason to think his architecture is optimal or especially well-supported by the evidence; however, it sounds at least a reasonable possibility. In fact, he tells us that certain aspects of his system have already worked well in real-life AI applications.

Unfortunately, I see a bigger problem. As I mentioned, the idea of scripts is not at all new. In earlier research they have delivered very good results when confined to a limited domain - ie when dealing with a smallish set of objects in a context which can be exhaustively described. Where they have never really succeeded to date is in producing the kind of general common sense which is characteristic of human cognition; the ability to go on making good decisions in changed or unprecedented circumstances, or in the seething ungraspable complexity of the real world. I see no reason to think that the schemata of Ouroboros are likely to prove any better at addressing these challenges.

Knud Thomsen very kindly sent the following response.

> *I feel honored by the inclusion of my tiny draft into this site!*
>
> *One of my points actually is that any account of mind and consciousness ought to be "quite practical and un-mystical". The second one, of course is, that ONE single grand overall PROCESS - account can do it all (rather than three stages: acquisition, evaluation, action...). This actually is the main argument in favor of the Ouroboros Model: it has something non-trivial to say about a vast variety of topics, which commonly are each addressed in separate models, which do not know anything about each other, not to mention that they together should form a coherent whole. The Ouroboros Model is meant to sketch an "all-encompassing" picture in a self-consistent, self-organizing and self-reflexive way.*
>
> *Proposals for technical details of neuronal implementation certainly explode the frame of this short comment; no doubt, re-entrant activity and biases in thalamocortical loops will play a role. Sure, emotional details and complexities are determined by the content of the situation; nevertheless, a most suitable underlying base could be formed by the feedback on how expectations come true. Previously established emotional tags would be one of the considered features and thus part of any evaluation, - "inherited", partly already from long-ago reptile ancestors.*
>
> *The Ouroboros Model offers a simple mechanism telling to what extent old scripts are applicable, what context seems the most adequate and when schemata have to be adapted flexibly and in what direction.*

Alien Consciousness

15 June 2008


I was reading somewhere about SETI and I was struck by the level of confidence the writer seemed to enjoy that we should be able, not only to recognise a signal from some remote aliens, but actually interpret it. It seems to me, on the contrary, that finding the signal is the 'easy problem' of alien communication. We might spend much longer trying to work out what they were saying than we did finding the message in the first place. In fact, it seems likely to me that we could never interpret such a signal at all.


Well, I don't think anyone underestimates the scope of the task, but you know it can hardly be impossible. For example, we send off a series of binary numbers; binary is so fundamental, yet so different from a random signal, that they would be bound to recognise it. The natural thing for them to do is echo the series back with another term added. Once we've got onto exchanging numbers, we send them, like say 640 and 480 over and over. If they're sophisticated enough to send radio signals, they're going to recognise we're trying to send them the dimensions of a 2d array. Or we could go straight to 3D, whatever. Then we can send the bits for a simple image. We might do that Mickey Mouse sort of picture of a water molecule: odds are they're water-based too, so they are bound to recognise it. We can get quite a conversation on chemistry going, then they can send us images of themselves, we can start matching streams of bits that mean words to streams of bits that mean objects, and we'll be able to read and understand what they're writing. OK, it'll take a long time, granted, because each signal is going to take years to be delivered. But it's eminently possible.


The thing is, you don't realise how many assumptions you're making. Maybe they never thought of atoms as billiard balls, and to them methane is more important than water anyway. Maybe they don't have vision and the idea of using arrays of bits to represent images makes no sense to them. Maybe they have computers that actually run on hexadecimal, or maybe digital computers never happened for them at all, because they discovered an analogue computing machine which is far better but unknown to us, so they've never heard of binary. But these are all trivial points. What makes you think their consciousness is in any way compatible with ours? There must be an arbitrarily large number of different ways of mapping reality; chances are, their way is just different to ours and radically untranslatable.


Every human brain is wired up uniquely, but we manage to communicate. Seriously, if there are aliens out there they are the products of biological evolution – no, they are, that's not just an assumption, it's a reasonable deduction – and that alone guarantees we can communicate with them, just as we can communicate with other species on Earth up to the full extent of their capability. The thing is, they may be mapping reality differently, but it's essentially the same reality they're mapping, and that means the two maps have to correspond. I might meet some people who use Urgh, Slarm, and Furp instead of North and South; they might use

cubits for Urgh distances, rods for Slarm ones, and holy hibbles for the magic third Furpian co-ordinate, but at the end of the day I can translate their map into one of mine. Because they are biological organisms, they're going to know about all those common fundamentals: death and birth, hunger, strife, growth and ageing, food and reproduction, kinship, travel, rest; and because we know they have technology they're going to know about mining and making things, electricity, machines – and communication. And that guarantees they 'll be able and willing to communicate with us.


You see, even on Earth, even with our fellow humans, it doesn't work. Look at all the ancient inscriptions we have no clue about. Ventris managed to crack Linear B only because it turned out to be a language he already knew; Linear A, forget about it. Or Meroitic. We have reams and reams of Meroitic writing, and the frustrating thing is, it uses Egyptian characters, so we actually know what it sounded like. You can stand in front of some long carved Meroitic screed and read it out, knowing it sounds more or less the way it was meant to; but we have no idea whatever what it means: and we probably never will.


What you're missing there is that the Cretans and the Meroitic people are never going to respond to our signals. The dialogue is half the point here: if we never get an answer, then obviously we're not going to communicate.

Though I like to think that even if we picked up the TV signal of some distant race which had actually perished long before, we'd still have a chance of working it out because there'd just be more of it, and a more varied sample than your Egyptian hieroglyphs which let's face it are probably all Royal memorials or something.


Look, you argue that consciousness is a product of evolution. But what survival advantage does phenomenal experience give us – how do qualia help us survive? They don't, because zombies could survive every bit as well without them. So why have we got them? It seems likely to me that we somehow got endowed with a faculty we don't even begin to understand. One by-product of this faculty was the ability to stand back and deliberate on our behaviour in a more detached way; another happened to be qualia, just an incidental free gift.


So you're saying aliens might be philosophical zombies? They might have no real inner experience? Somehow I knew it would come back to qualia eventually.


More than that - I'm not just saying they might be zombies, but that's one possibility, isn't it? Incidentally, I wonder what moral duty would we owe to creatures that had no real feelings? Would it matter how we behaved towards them…? Curious thought.


Even zombies have rights, surely? Whatever the ethical theory, we can never be sure whether they actually are zombies, so you'd have to treat them as if they had feelings, just to be on the safe side. Anyway, to be honest, I don't think I like where you seem to be going with this.

 No, well the point is not that they might be zombies, but that instead of our faculty of consciousness, they might have a completely different one which nevertheless served the same purpose from an evolutionary point of view and had similar survival value. We're the first animals on Earth to evolve a consciousness: it's as if we were primitive sea creatures and have just developed a water squirt facility, making us able to move about in a way no other creature can yet do. But these aliens might have fins. They might have legs. You sit there blandly assuming that any mobile creature will want to match squirts with us, but it ain't necessarily so.

 No, your analogy is incorrectly applied. I'm not saying they'd want to match squirts; I'm saying we'd be able to follow each other, or have races, or generally share the commonalities of motion irrespective of the different kit we might be using to achieve that mobility.

 Your problem here is really that you can't imagine how an alien could have something that delivered the cognitive benefits of consciousness without being consciousness itself. Of course you can't; that's just another aspect of the problem; our consciousness conditions our view of reality so strongly that we're not capable of realising its limitations.

 Look, if we launch a missile at these people, they'll send us a signal, and that signal will mean Stop. It's those cognitive benefits you dismiss so lightly that I rest my case on. For practical purposes, about practical issues, we'll be able to communicate; if they have extra special 3d rainbow qualia, or none, or some kind of ineffable quidlia that we can never comprehend – I don't care. You might have alien quidlia buzzing round your head for all I know; that possibility doesn't stop us communicating. Up to a point.

 You know that Wittgenstein said that if a lion could speak, we couldn't understand what he was saying?

 Yes, I do know; just one of many occasions when he was talking balls. In fairness to old Wittless he also said 'If I see someone writhing in pain with evident cause I do not think; all the same, his feelings are hidden from me.'

And the same goes for writhing aliens.

Language and Consciousness

29 June 2008

There is clearly a close relationship between consciousness and language. The ability to conduct a conversation is commonly taken as the litmus test of human-style consciousness for both computers and chimpanzees, for example. While the absence of language doesn't prove the absence of consciousness - not all of our thoughts are in words - the lack of a linguistic capacity seems to close off certain kinds of explicit reflection which form an important part of human cognition. Someone who had no language at all might be conscious, but would they be conscious in quite the same way as a normal, word-mongering human? It might therefore seem that when Jordan Zlatev asserts the dependence of language on consciousness, he is saying something uncontroversial. In fact, he has both broader and more specific aims: he wants to draw more attention to the relationship on the one hand, and on the other readjust our view of where the borders between conscious and unconscious processes lie.

It seems pretty clear that a lot of the work the brain does on language is unconscious. When I'm talking, I don't, for example, have to think to myself about the grammar I'm using (unless perhaps I'm attempting a foreign language, or talking about some grammatical point). I don't even know how my brain operates English grammar; it surely doesn't use the kind of rules I was taught at school; perhaps in some way it puts together Chomskyan structures, or perhaps it has some altogether different approach which yields grammatical sentences without anything we would recognise as explicit grammatical rules. Whatever it does, the process by which sentences are formed is quite invisible to me, the core entity to whom those same sentences belong, and whose sentiments they communicate. It seems natural to suppose that the structure of our language is provided pre-consciously.

Zlatev, however, contends that the rules of language are social and normative; to apply them we have to understand a number of conventions about meanings and usage; and whatever view we may take of such conventions their use requires a reflective knower (Zlatev picks up on a distinction set out by Honderich between affective, perceptual, and reflective consciousness; it's the latter he is concerned with). To put it another way, operating the rules of language requires us to make judgements of a kind which only reflection can supply, and reflection of this kind deserves recognition as conscious. Zlatev is not asserting that the rules of grammar at work in our brain are consciously know after all: he draws a distinction between accessibility and introspectability; he wants to say that the rules are known pre-theoretically, but not unconsciously.

Perhaps we could put Zlatev's point a different way: if the rules of language were really unconscious, we should be incapable of choosing to speak ungrammatically, just as we are incapable of making our heart beat slower or our skin stop sweating by an act of will. Utterances which did not follow the rules would be incomprehensible to us. In fact, we can cheerfully utter malformed sentences, distinguish them from good ones and usually understand both. Deliberate transgressions of the rules are used for communicative or humorous effect (a rather Gricean point). While the theory may be hidden from introspection, the rules are accessible to conscious thought.

If the rules of language were unconscious, asks Zlatev, how would we account for the phenomenon of self-correction, in which we make a mistake, notice it, and produce an emended version? And how could it be that the form of our utterances is often structured to enhance the meaning and distribute the emphasis in the most helpful way? An unconscious sentence factory could never support our conscious intentions in such a nicely judged way. Zlatev also brings forward evidence from language acquisition studies to support his claim that unconscious mechanisms may support, but do not exhaust, the processes of language.

At times Zlatev seems to lean towards a kind of Brentano-ish view; language requires intentionality, and nothing but consciousness can provide it (alas, in a way which remains unexplained). Intriguingly, he says that he and others were deceived into accepting the unconsciousness of language production at an earlier stage by the allure of connectionism, whose mechanistic nature only gradually became clear. I think connectionists might feel this is a little unfair, and that Zlatev need not have given up on connectionist approaches to reflective judgement simply because they are 'mechanistic'.

All in all, I think Zlatev offers some useful insights; his general point that a binary division between conscious and unconscious simply isn't good enough is indeed a good one and well made. I wonder whether this is a point particular to the language faculty, however. Couldn't I make some similar points about my tennis-playing faculty? Here too I rely on some unconscious mechanisms and couldn't tell you exactly which arm muscles I used in which way. Yet making my hand twist the racquet around and move it to the right place also seems to require some calculated, dare I say reflective, judgements and the way I do it is exquisitely conditioned by tactics and strategy which I devise and entertain at a fully self-conscious level.

Be that as it may, it's bad news for the designers of translation software if Zlatev is right, since their systems will have to achieve real consciousness before they can be perfected.

Only Integrate

19 July 2008

Consciousness does not require language. Nor does it require memory. Or perception of the world, or any action in the world, or emotions, or even attention. So say Christof Koch and Giulio Tononi.

(Koch and Tononi? Sounds an interesting collaboration - besides their own achievements both in the past have been co-authors of grand panjandrums of the field - Koch worked with Francis Crick and Tononi with Gerald Edelman. They have some new ideas here, however.)

I don't think everyone would agree that consciousness can be stripped down quite as radically as Koch and Tononi suggest. I actually find it rather tricky to get a good intuitive grasp of exactly what it is that's left when they've finished: it seems to be a sort of contentless core of subjectivity. Particular items on the list of exclusions also seem debatable. We know that some unfortunate people are unable to create new memories, for example, but even in their case, they retain knowledge of what's going on long enough to have reasonable short-term conversations. Could consciousness exist at all if the contents of the mind slid away instantly? Or take action and perception; some would argue that consciousness requires embodiment in a dynamic world; some indeed would argue that consciousness is irreducibly social in character. Again we know that some unlucky people have been isolated from the world by the loss of their senses and their ability to move, without thereby losing consciousness: but these are all people who had perception and action to begin with. Actually Koch and Tononi allow for that point, saying that whether consciousness requires us to have had certain faculties at some stage is another question; they merely assert that ongoing present consciousness doesn't require them.

The most surprising of their denials is the denial of attention - some people would come close to saying that consciousness was attention. If there were no perception, no memory , and no attention one begins to wonder how consciousness could have any contents: they couldn't arrive via the senses; they couldn't be called up from past experience; and they couldn't even be generated by concentrating on oneself or the surrounding nullity.

However, let's not quibble. Consciousness has many meanings and covers a number of related but separable processes. The philosophical convention is that he who propounds the thesis gets to dictate the definitions, so if Koch and Tononi want to discuss a particularly minimal form, they're fully entitled to do so.

What, then, do they offer as an alternative essence of consciousness? In two words, integrated information. A conscious mind will require many - hugely many - possible states; but, they argue, this is quite different from the case of a digital camera. A 1 megapixel camera has lots of possible states (one for every frame of every possible movie, they suggest - not sure that's strictly true since some frames will show identical pictures, not that that makes a real

difference); but these states are made up of lots of photodiodes whose states are all fully independent, By contrast, states of the mind are indivisible; you can't choose to see colours without shapes, or experience only the left side of your visual field.

This sounds like an argument that consciousness is not digital but irreducibly analogue, though Koch and Tononi don't draw that conclusion explicitly, and I don't think either of them plan to put away their computer simulations. We can, of course, draw a distinction between consciousness itself and the neural mechanisms that support it, so it could be that integrated, analogue conscious experience could be generated by processes which themselves are digital or at least digitizable.

At any rate, they call their theory the Integrated Information Theory (IIT); it suggests, they say, a way of assessing consciousness in a machine, a superior Turing Test (they seem to think the original has been overtaken by the performance of the chatterbot Alice - with due respect to Alice, that seems premature, though I have not personally tried out Alice's Silver Edition). Their idea is that the candidate should be shown a picture and asked to provide a concise description; human beings will easily recognise, for example, that the picture shows an attempted robbery in a shop and many other details, whereas machines will struggle to provide even the most naive and physical descriptions of what is depicted. This would surely be an effective test, but why does integrated information provide the human facility at work here? It's a key point about the theory that as well as elements of current experience being integrated into an indivisible whole, current experience is integrated with a whole range of background information: so the human mind can instantly bring into play a lot of knowledge the machine can't access.

That's all very well, but this concept of integration is beginning to look less like a simple mechanism and more like a magic trick. Koch and Tononi offer a measure of integrated information, which they call phi: it represents the reduction in uncertainty, expressed in bits, when a system enters a particular state. Interestingly, the idea is that high values of phi correspond with higher levels of consciousness on a continuous scale; so animals are somewhat less conscious than us on average; but it also must be possible to be more conscious than any human to date has ever been: in fact, there is no upper limit in theory to how conscious an entity might be. This is heady stuff which could easily lead on to theology if we're not careful.

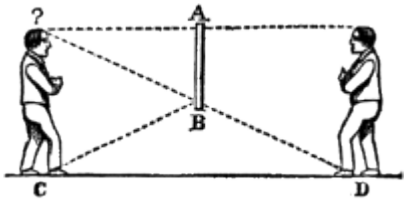
To illustrate this idea of the reduction of uncertainty, the authors compare our old friend the photodiode with a human being (suppose it to be a photodiode with only two states), When the lights go out, the photodiode eliminates the possibility of White, and achieves the certainty of Black. But for the human being, darkness eliminates a huge range of possibilities that you might have been seeing; a red screen, a blue one, a picture of the statue of Liberty, a picture of your child's piano recital, and so on. So the value of phi, the reduction in uncertainty, for human beings is much vaster than that for the photodiode. Of course measuring phi is not going to be straightforward; Koch and Tononi say that for the moment it can only be done for very simple systems. That's fair enough, but one aspect of the argument seems problematic to me. The reduction of uncertainty on this argument seems to be, not vast, but infinite. When the lights went out in your competition with the photodiode, you could have been seeing a picture of one grain of sand; you could have been seeing a picture of two grains of sand, and so on for ever. So it seems the value of phi for all conscious states is infinite. Hang on, though; is even darkness that simple? Remembering that Koch and Tononi's integration is not limited to current experience, the darkness might cause you to think of what it's like when you're about to fall

asleep; an unexpected hiatus in the cinema eighteen months ago: a black cat in a coal cellar at midnight. In fact, it might seem like the darkness which enshrouds a single grain of sand; the indefinably bulkier darkness which hides two grains of sand… So don't both states involve an infinite amount of information, so that the value of phi for all states of consciousness is zero?

I think Koch and Tononi have succeeded in targeting some of the real fundamental problems of consciousness, and made a commendably bold and original attack on them. But they will need a very clear explanation of how their kind of integration works if IIT is really going to fly.

Why Am I Me?

10 August 2008



There are a few philosophical problems which occur spontaneously to people who know nothing of academic philosophy but have a naturally thoughtful inclination. The problem of free will is one, I think, and probably so is qualia; many people who never heard of David Chalmers sometimes ponder the 'hard question', asking themselves how they know that the blue they see 'in their heads' is the same as the blue other people see. David M. Black has put his finger on another of these problems in his paper on The Ownership of Consciousness. Why am I me and not someone else? Black's main purpose lies elsewhere - he wants to suggest that talk of spirituality can be a valuable way of discussing structures in the subjective part of the world, complementing the reductive scientific account which deals with the objective aspects. That's an attractive project (though I think he allows himself too much too easily in assuming that subjective experience has causal effects).But it was the issue posed by the title of the paper that particularly caught my attention.

Now of course, there is a sense in which this is is an absurd enquiry. Whoever I am, that person is me: I can't not be me, by definition. Self and consciousness are intertwined, so that the ownership of my consciousness can never really be in doubt. I am my consciousness. It would make more sense to ask why I have this body than to ask why I have this consciousness. So it might seem that the mystery of why I am who I am is really about on a par with the mystery of how I was lucky enough to be born on Earth, rather than on a planet without an atmosphere; not really a mystery at all.

But suppose, we might say, we strip away the details of my body and my life and pare me down to the essential nub of experiencing entity. What makes this nub any different from other such nubs, and why is it linked with the life of this particular human organism rather than any other?

Some would say in response that there is no such nub; it's exactly my history and my physical constitution that make my consciousness what it is; so again it's no surprise, properly understood, that my experiences belong to me and not to anyone else. Strip away all those supposedly inessential features, and you strip me away with them. Others would accept that it's my history and composition that define me as me, but feel, possibly on the basis of introspection, that there still is something to me over and above all that. They might find this final ingredient in an inscrutable panpsychic quality of matter itself; others have suggested a kind of universal experiential substrate or background. Instead of individual nubs, we have a kind of Universal self; a line of thought which is highly compatible with some religious and mystical views.

However, I think both sides of this argument are missing the point. To say that my individual consciousness arises from or is constituted by my physical nature and background is not actually to dispose of the essential problem at all, because my physical nature and background are also inexplicably particular. Perhaps the underlying problem is the vexed question of why anything is anything in particular: it's just that in the case of my own experience and my own existence the question hits me with a force it lacks when I'm merely wondering about a chair. The basic problem is haecceity or thisness (the same problem which in my view lies at the root of the qualia problem).

The difficulty of this issue is that it seems no kind of explanation will do. One way to explain the arbitrary complexity of the world would be to assume that everything is really a logical necessity, so if we were clever enough we could deduce all truths from first principles, like Douglas Adams' Deep Thought which, beginning with the cogito had got as far as deducing the existence of rice pudding and income tax before anyone could switch it off, if I remember correctly. Even if we could pull off such staggering feats of deduction, the explanation is no good because if everything exists only by virtue of logical necessity everything exists in an eternal Platonic world, the opposite of the mutable diversity we set out to explain. A more scientific view would have it that the current state of the world derives from previous states in accordance with the laws of physics, so that the explanation is essentially historical. But this is no better. We now have an expanded set of laws; beside those of logic, we need those of mathematics and those of physics. But they're still laws, so we're still Platonic and immutable unless some arbitrary graininess somehow crept in at the beginning of it all. Nowadays we're readier to accept that huge chaotic complexity can arise out of small beginnings; but not out of nothing. Explaining that original graininess is as difficult as explaining the haecceity of the world was to start with.

And what about those laws of physics? Do they too reduce to logical or mathematical necessity, or is there some arbitrary element involved; and if so, how do you explain that? One line we could take is to say that all the possible variations of the underlying constants are realized in some possible world. The problem of how we got these particular laws of physics then turns into the same sort of vacuous problem as the ones mentioned above: if the laws weren't like that, you wouldn't be here to wonder about it.

But that's no good. Even if we could get over the problems involved in the idea of parallel worlds, which seems to involve the problematic retention of identity between non-identical entities, how can we deal with the concept of all possible sets of laws of physics? Typically in these discussions it is assumed that we're talking about variations in the value of a few constants, but things are much worse than that. Apart from the sheer bogglement of universes where the value of gravity is determined by quadripedal blurpton interactions in the trouser-pocket of fourth-order fried bread, if all possible sets of laws are realised, that includes universes whose laws and constitution are the same as ours up until 2009, when the blurptons abruptly take over. In short, anything could happen at any time; to say that all possible laws of physics are realised in some universe is in effect to declare that there are no laws of physics and that everything happens arbitrarily.

That is another possible position, of course, though an utterly unsatisfactory one. Taking a more traditional tack, we could say that the world is the way it is because of the will of God; but for philosophers, that's no good. We need to know whether God was working on logical principles, and if it wasn't pure logic, where did His axioms or His quirks of personality come from?

In the last century it was finally established that there is something in maths that isn't reducible to logic; not much, but an essential little something which can be construed in different ways. It seems to me that in the same way there's some fundamental element in metaphysics that isn't any kind of law; but I have no idea how to construe it at all.

Feelings about Jaynes

7 September 2008

I see that the annual Julian Jaynes Conference took place last month. As you may know, Jaynes put forward a surprising theory of consciousness which suggested it had a relatively recent origin. According to Jaynes ancient human beings, right up into early historical times, had minds that were divided into two chambers. One of these chambers was in charge of day-to-day life, operating on a simple, short-term emotional basis for the most part (though still capable of turning out some substantial pieces of art and literature, it seems). The occasional interventions of the second chamber, the part which dealt in more reflective, longer-term consideration were not experienced as the person's own thoughts, but rather as divine or ancestral voices restraining or instructing the hearer, which explains why interventionist gods feature so strongly in early literature. The breakdown of this bicameral arrangement and the unification of the two chambers of the mind were, according to Jaynes, what produced consciousness as we now understand it.

I find this bicameral theory impossible to believe, but it does have some undeniably engaging qualities. The way it gives a neat potential explanation for divine voices and for certain modern mental disorders gives it a superficial plausibility, especially when expounded with Jaynes' characteristic eloquence and panache. It's tempting to think of it as a drastically overstated version of an essentially sound insight, but even if it's completely wrong, thinking about its flaws is a stimulating exercise.

At this year's conference, Stevan Harnad gave a speech in two parts, beginning with some personal reminiscences of Jaynes - he mentions that he found it impossible to write an obituary for Jaynes and you get the feeling that his emotions are still making it a difficult subject for him to talk about - and then going on to a philosophical discussion of the interesting question of whether Jaynes would have kicked a dog, and why not.

Why shouldn't he? For Jaynes, after all, consciousness was uniquely human; no other creature had gone through the breakdown of a bicameral mind. There's nothing especially wrong with kicking unconscious objects and since dogs lack consciousness, there should be no particular reason not to kick one; but Harnad was sure Jaynes, a gentle and civilised man, would certainly not have done so. In fact, he had confirmed this in conversation with Jaynes during his lifetime. Jaynes said it was true that dogs in themselves did not deserve the same moral consideration as conscious entities like human beings; but that we should by all means refrain from kicking dogs unnecessarily because of the moral consequences for the kicker and onlookers. Kicking dogs is a bad, desensitising act, unworthy of the moral dignity of human beings even though dogs don't fundamentally matter.

This is an interesting answer, and I think it's intellectually tenable. We should, on the whole, refrain from slashing rose bushes to pieces, and from smashing beautiful porcelain, even though plants and pots are not conscious. But as Harnad suggests, we may doubt the sincerity of the argument – it has the air of a rationalisation run up to defend a weak spot in a wider case rather than something sincerely believed. We might think that a better line of argument was

available to Jaynes if he had been willing to say that unconscious creatures can still be moral objects, which is surely true.

Harnad, ultimately, wants to say that dogs are indeed moral objects because they have feelings, or so our mirror neurons tell us; and that empathy is enough to make us hold off, even Julian Jaynes. Although I suspect this overstates the role of mirror neurons, he's surely right to think that the possession of feelings is enough to constitute a moral object. As Jeremy Bentham put it, 'The question is not, can they reason?, nor can they talk? but, can they suffer?'

What's particularly interesting is the discussion Harnad provides about feelings (in the loosest sense; any kind of mental intimation, including but not limited to sensory input). He begins by pointing out that Descartes' cogito ergo sum is not a logical deduction but a claim about the infallibility of certain thoughts or feelings. To think that one exists can't be a mistake because non-existent people don't think at all. Harnad scrupulously points out that it's the existence of the the thought itself which is established, rather than the existence of the more problematic self. However, other feelings have a similar kind of infallibility; we can be wrong about whether we've got a bad tooth, but not that it feels like toothache. Harnad notes that a similar kind of infallibility attaches to what we mean or understand. We can of course use words that don't, in the wider world, have the meaning we wanted, but we can't be wrong about what we meant internally. Harnad describes this as the distinction between wide and narrow meaning; it largely corresponds with the more controversial distinction between intrinsic and derive intentionality (thoughts have meanings because somehow they just do; books have meanings only because they record and evoke thoughts).

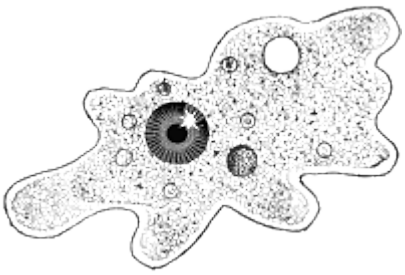
It all comes down to feelings according to Harnad. "2x2=4" does not feel the same to us as "Kétszer kettő négy", but if, like him, we spoke Hungarian, they would feel very similar, because they mean the same thing.

This is very interesting. The problem of intentionality, of meaningfulness, is one of the principal problems of consciousness, but it tends in my view to be somewhat neglected - perhaps partly because it's so difficult to find anything worth saying about it. New ideas in this area are very welcome, and on the face of it Harnad's suggestion is plausible (sincerity and strong feelings seem to go together at any rate). The chief problem, perhaps, is that even if it's true, this insight doesn't move us on as much as we should like. There's no accompanying theory of feelings, and since Harnad has explicitly chosen the vaguest and widest interpretation of the term, we still don't know all that much about the fundamental nature of intentionality.

My feeling, on the whole, is that in fact the true essence of meaningfulness lies elsewhere; a feeling that x is an invariable accompaniment to believing that x, but does not constitute the belief. Two cheers for Harnad, though, and a third for Jaynes, whose legacy remains so productive.

Nanointentionality

20 September 2008



Somewhat belatedly I came across an interesting paper by W Tecumseh Fitch the other day (Actually I came across Beau Siever's discussion of Daniel Dennett's discussion of the paper.) in which he boldly tackles the thorny subject of original intentionality, claiming it's all based on what he calls nano-intentionality.

Fitch declares himself a defender of intrinsic intentionality. Intentionality, as you may know, is aboutness, meaningfulness: things like books and films are said to have derived intentionality: they are about things because the people who made them and the people who read or watch them interpret them as being about something, conferring meaning on them. But some things, our own thoughts, for example, are not about things because of what anyone else thinks, they just are intrinsically about things. How they manage this has always been a mystery.

Dennett, in fact, denies that there is any such thing as intrinsic intentionality - how can anything just inherently be about something? It's this view that Fitch wants to challenge; strangely, Dennett says it's all a misunderstanding and he agrees with Fitch.

How can this be? Well, Dennett would be right to reject intrinsic intentionality if it meant that we just say things are magically about things and that's the end of the story; but really when people speak of intrinsic intentionality it is usually a kind of promissory note: they mean, here in people's minds is where meaning originates; we don't know how yet, but meanings in minds are different to meanings in books. Fitch means to say; this is where meaning originates, and I think I can tell you how. I think Dennett is comfortable with theories of intentionality which provide a decent naturalistic interpretation - and if that's what we're doing, he doesn't really care too much whether we're calling it intrinsic, or original, or whatever.

Fitch's view still seems at odds with Dennett's in many ways; he rejects the idea that computers could have intrinsic intentionality, and in general his ideas would seem to fit well with those of Searle, Dennett's archenemy. Searle says that consciousness arises from certain kinds of biological material as a result of some properties of that material which we don't understand (yet - Searle is sure that further scientific research will enable us to understand them). Nanointentionality would seem to fit into that view quite well.

So what is it? Fitch says that biological organisms exhibit a responsiveness to their environment which no machine can emulate. When we're cut, we heal up: our flesh extemporises, forming functional but ad hoc patterns of tissue that patch up the gap. Amoebas and smaller single-celled organisms respond to their surroundings, not just in a pre-organised way, but flexibly, managing to respond and adapt even to new circumstances. This kind of responsiveness to the environment, in his view, is the elementary precursor to true intentionality: the responses are not, in detail at least, writtn into the organism, and they are, at a basic level, goal-directed.

Having, as he believes, obtained this narrow foothold, Fitch seeks to build on it. Cells working together can build up an information processing system which inherits from them the spark of aboutness while adding to it new capacities. When they reach the level where options can be modelled and accepted or rejected, full-blown consciousness and true intrinsic intentionality

dawn. There's something a little surprising about this; Fitch is relying for most of the work on the kind of functional organisation he otherwise rejects. At least half of the powers of intentionality seem to come, not from the initial spark, but the way the neurons process information - the sort of thing you might think was perfectly amenable to computation (I see Dennett nodding happily). It prompts another thought, too: Fitch denies that mere silicon can have the kind of open-ended responsiveness of an amoeba. If we swap cells for transistors, that may be right; but what if the computer moves down a notch and simulates the parts (perhaps even the molecules) of the amoeba? Since Fitch is committed to naturalism, it seems hard to exclude computers from having the properties of living things so absolutely as he wishes.

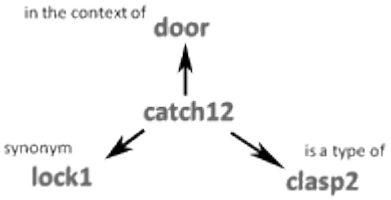
There is also a problem down there with the nano-intentionality, too. Fitch sees the responsiveness of the eukaryotic organisms (he's prepared to exclude the prokaryotes) as having a directedness which prefigures proper intentionality. But I doubt it. This directedness is a real and interesting quality, resembling what Grice called natural meaning: those spots mean measles; that spade means a hole to be dug. This is a good place to be looking for the roots of intentionality; indeed, my own view would have them somewhere in this area. The snag is that natural meaning has a tendency to separate out into two parts; the fitness of a thing for a result, and derived intentionality. The first of these has nothing of intentionality in it, properly understood; a spade may be specially fit for digging, but a large snowbank is especially fit for avalanches; large black clouds are fit for rain; there's no real meaning at work. The element of derived intentionality lies in the eye of the beholder; the spade is about digging because that's the way it was designed, and that's the way I mean to use it. This is intentionality, but resoundingly not the intrinsic kind we're after.

So, if we look at Fitch's amoeba, we can analyse its responsiveness. In part it's simply a fitness to go on surviving no different in principle from the cloud's fitness to rain; in part it's a purposefulness which we and Fitch can't help reading into it. Take away these two elements, and the foothold on which Fitch stands disappears, I think.

Cognition Incorporated

4 October 2008


So, another bastion of mysterianism falls. Cognition Technologies Inc has a semantic map of virtually the entire English language, which enables applications to understand English words. Yes, understand. It's always been one of your bedrock claims that computers can't really understand anything; but you're going to have to give that one up.




Oh, yes, I read about that. In fact it's been mentioned in several places. I thought the Engadget headline got to the root of the issue pretty well - 'semantic map opens way for robot uprising'. I suppose that's pretty much your view. It seems to me just another case of the baseless hyperbole that afflicts the whole AI field. There are those people at Cyc and elsewhere who think cognition is just a matter of having a big enough encyclopaedia: now we've got this lot who think it's just a colossal dictionary. But having a long list of synonyms and categories doesn't constitute understanding; or my bookshelf would have achieved consciousness long ago.


Let me ask you a question. Suppose you were a teacher? How would you judge whether your pupils understood a particular word? Wouldn't you see whether they could give synonyms, either words or phrases that meant the same thing? If you yourself didn't understand a word what would you do? You'd go over to that spooky bookcase and look at a list of synonyms. Once you'd matched up your new word with one or two synonyms, you'd understand it, wouldn't you?

I know you've got all these reservations about whether computers really this or really that. You don't accept that they really learn anything, because true human learning implies understanding, and they haven't got that. They don't really communicate, they just transfer digital blips. According to you all these usages are just misleading metaphors. According to you they don't even play chess, not in the sense that humans do. Now frankly, it seems to me that when the machine is sitting there and moving the pieces in a way that puts you into checkmate, any distinction between what it's doing and playing chess is patently nugatory. You might as well say you don't really ride a bike because a bike hasn't got legs, and that bike-riding is just a metaphor. However, I recognise that I'm not going to drive you out of your cave on this. But you're going to have to accept that machines which use this kind of semantic mapping can understand words at least to the same extent that computers can play chess. Concede gracefully; that'll be enough for today.


I'll grant you that the mapping is an interesting piece of work. But what does it really add? These people are using the map for a search engine, but is it really any better than old-fashioned approaches? So we search for, say 'tears'; the search engine turns up thousands of pages on weeping, when we wanted to know about tears in a garment. Now Cognition's engine will be able to spot from the context what I'm really after. Because I've searched for 'What to do when something tears your trousers' it will notice the word trousers and give me results about

rips. But so will Google! If I give Google the word trousers as well as tears, it will find relevant stuff without needing any explicit semantics at all. These people don't understand why a 'semantic map' is necessary for a search engine, and they're sceptical about Cognition's ontology (in that irritating non-philosophical sense).

 Wrong! When you do a search on Google for 'What to do when something tears your trousers' the top result is actually about tears from your eye, believe it or not. Typically you get all sorts of irrelevant stuff along with a few good results. But to see the point, look at an example Cognition give themselves. Their legal search demo (try it for yourself), when asked for results relating to 'fatal fumes in the workplace' came up with a relevant result which contains neither 'fatal' nor 'workplace', only 'died' and 'working'. This kind of thing is going to be what Web 3 is built around.

 If this thing is so good, why haven't they used it for a chatbot? If this technology involves complete understanding of English, they ought to breeze through the Turing test. I'll tell you why: because that would involve actual understanding, not just a dictionary. Their machine might be able to look up meaning of pitbull and find that the synonym dog, but it wouldn't have a hope with the lipstick pitbull, or are they the ones with all the issues.

 Nobody says the Cognition technology has achieved consciousness.

 I think you are saying that, by implication. I don't see how there can be real understanding without consciousness.

 Or if there is, it won't be real consciousness according to you - just a metaphor...

Loebner 2008

14 October 2008

The annual Loebner Prize has been won by Elbot. As you may know, the Loebner competition implements the Turing Test, inviting contestants to put forward a chat-bot program which can conduct online conversation indistinguishable from one conducted with a human being. We previously discussed the victory of Rollo Carpenter's Jabberwacky, a contender again this year.

One of Elbot's characteristics, which presumably helped tip the balance this year, is a particular assertiveness about trying to manage the conversation into 'safe' areas. One of the easiest ways to unmask a chat-bot is to exploit its lack of knowledge about the real world; but if the bot can keep the conversation in domains it is well-informed about, it stands a much better chance of being plausible. Otherwise the only option is often to resort to relatively weak default responses ('I don't know', 'What do you think?', 'Why do you mention [noun extracted from the input sentence]?).

But aren't Elbot's tactics cheating? Don't these cheap tricks invalidate the whole thing as a serious project? Some would say so: the Loebner does not enjoy universal esteem among academics, and Marvin Minsky famously offered a cash reward to anyone who could stop the contest.

We have to remember, however, that the contestants are not seeking to reproduce the real operation of the human brain. Humans are able to conduct general conversation because they have general-purpose consciousness, but that is far too much to expect of a simple chat-bot. The Turing Test is sometimes interpreted as a test for consciousness, but that isn't quite how Turing himself described it (he proposed it as a more practical alternative to considering the question 'Can machines think?').

OK so it's not cheating, but all the same, if it's just fakery, what's the value of the exercise? There are several answers to this. One is the 'plane wing' argument: planes don't fly the way birds do, but they're still of value. It might well be that a program that does conversation is useful in its own right, even if it doesn't do things the way the human brain does; perhaps for human/machine interfaces. On the other hand, as a second answer, it might turn out that discoveries we make while making chat-bots will eventually shed some light on how some parts of the brain put together well-structured and relevant sentences. If they don't do that, they may still lead to the discovery of unexpectedly valuable techniques in programming: solving difficult but apparently pointless problems just for the hell of it does sometimes prove more fruitful than expected. A fourth point which I think perhaps deserves more attention is that even if chat-bots tell us nothing about AI, they may still tell us interesting things about human beings. The way trust and belief are evoked, for example: the implicit rules of conversation, and the pragmatics of human communication.

The clincher in my view, however, is that the Loebner is fun, and enlivens the subject in a way which must surely be helpful overall. How many serious scientists were inspired in part by a secret childhood desire to have a robot friend they could talk to?

In a way you could say Elbot is attempting to refine the terms of the test. A successful program actually needs to deploy several different kinds of ability, and one of the most challenging is bringing to bear a fund of implicit background knowledge. No existing program is remotely as good at this as the human brain, and there are some reasons to doubt whether they ever will be. In the meantime, at any rate, there may be an argument for taking this factor out of the equation: Elbot tries to do this by managing the conversation, but in some early Loebner contests the conversations were explicitly limited to particular topic, and maybe this approach has something to be said for it. I believe Daniel Dennett, who was once a Loebner judge, suggested that the contest should develop towards testing a range of narrower abilities rather than the total conversational package. Perhaps we can imagine tests of parsing, of knowledge management, and so on.

At any rate, the Loebner seems in vigorous health, with a strong group of contenders this year: I hope it continues. [Note: It did, but only until 2019. With the arrival of programs using LLMs the Turing Test ceased to be credible. We now had AI that could turn out extremely convincing prose on any subject but did not even claim to have any kind of consciousness]

Consciousness: the new battleground for Creationists?

25 October 2008

A piece in the New Scientist suggests that creationists and their sympathisers are seeking to open up a new front. They think that the apparent insolubility of the problem of qualia means that materialism is on the way out; in fact, that consciousness is 'Darwinism's grave'. Cartesian dualism is back with a vengeance. Oh boy: if there was one thing the qualia debate didn't need, it was a large-scale theological intervention. Dan Dennett must be feeling rather the way Guy Crouchback felt when he heard about the Nazi-Soviet pact: the forces of darkness have drawn together and the enemy stands clear at last!

The suggested linkage between qualia and evolution seems tortuous. The first step, I suppose, assumes that dualism makes the problem of qualia easier to solve; then presumably we deduce that if dualism is true, it might as well be a dualism with spirits in (there are plenty of varieties without; in fact if I were to put down a list of the dualisms which seem to me most clear and plausible, I'm not sure that the Christian spirit variety would scrape into the Top Ten); then, that if there are spirits, there could well be God, and then that if there's God he might as well take on the job of governing speciation. At least, that's how I assume it goes. A key point seems to be the existence of some form of spiritual causation. Experiments are adduced in which the subjects were asked to change the pattern of their thoughts, which was then shown to correspond with change in the activity of their brain; this, it is claimed, shows that mind and brain are distinct. Unfortunately it palpably doesn't; attempting to disprove the identity of mind and brain by citing a correlation between the activity of the two is, well, pretty silly. Of course the thing that draws all this together and makes it seem to make sense in the minds of its advocates is Christianity, or at any rate an old-fashioned, literalist kind of Christianity.

Anyway, I shall leave Darwinism to look after itself, but in a helpful spirit let me offer these new qualophiles two reasons why dualism is No Good.

The first, widely recognised, is that arranging linkages between the two worlds, or two kinds of stuff required by dualism, always proves impossible. In resurrecting 'Cartesian dualism' I don't suppose the new qualophiles intend to restore the pineal gland to the role Descartes gave it as the unique locus of interaction between body and soul, but they will find that coming up with anything better is surprisingly difficult. There is a philosophical reason for this. If you have causal connections between your worlds - between spirits and matter, in this case - it becomes increasingly difficult to see why the second world should be regarded as truly separate at all, and your theory turns into a monism before your eyes. But if you don't have causal connections, your second world becomes irrelevant and unknowable. The usual Christian approach to this problem is to go for a kind of Sunday-best causal connection, one that doesn't show up in the everyday world, but lurks in special invisible places in the human brain. This was never a very attractive line of thinking and in spite of the quixotic efforts of those two distinguished knights, John Eccles and Karl Popper, it is less plausible now than ever before, and its credibility drains further with every advance in neurology.

The second problem, worse in my view, is that dualism doesn't really help. The new qualophile case must be, I suppose, that our ineffable subjective experiences are experiences of the spirit,

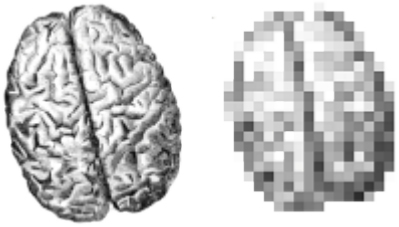
and that's what gives them their vivid character. The problem of qualia is to define what it is in the experience of seeing red which is over and above the simple physical account; bingo! It's the spirit. To put it another way, on this view zombies don't have souls.

But why not? How does the intervention of a soul turn the ditchwater of physics into the glowing wine of qualia? It seems to me I could quite well imagine a person who had a fully functioning soul and yet had no real phenomenal experiences: or at any rate, it's as easy to imagine that as an unsouled zombie in the same position. I think the new qualophiles might reply that my saying that shows I just haven't grasped what a soul is. Indeed I haven't, and I need them to explain how it works before I can see what advantage there is in their theory. If we're going to solve the mystery of qualia by attributing it to 'souls', and then we declare 'souls' a mystery, why are we even bothering? But here, as elsewhere with theological arguments, it seems to be assumed that if we can get the question into the spiritual realm, the enquiry politely ceases and we avert our eyes.

It is, of course, the same thing over on the other front, where creationists typically offer criticism of evolutionary theory, but offer not so much as a sniff of a Theory of Creation. Perhaps in the end the whole dispute is not so much a clash between two rival theories as a dispute over whether we should have rational theories at all.

Whole Brain Emulation

8 November 2008



Robots.net recently featured the Whole Brain Emulation Roadmap produced by the Future of Humanity Institute at Oxford University. The Future of Humanity Institute has a transhumanist tinge which I find slightly off-putting, and it does seem to include fiction among its inspirations, but the Roadmap is a thorough and serious piece of work, setting out in summary at least the issues that would need to be addressed in building a computer simulation of an entire human brain. Curiously, it does not include any explicit consideration of the Blue Brain project, even in an appendix on past work in the area, although three papers by Markram, including one describing the project, are cited.

One interesting question considered is: how low do you go? How much detail does a simulation need to have? Is it good enough to model brain modules (whatever they might be), neuronal groups of one kind of another, neurons themselves, neurotransmitters, quantum interactions in microtubules? The roadmap introduces the useful idea of scale separation; there might be one or more levels where there is a cut-off, and a simulation in terms of higher level entities does not need to be analysed any further. Your car depends on interactions at a molecular level, but in order to understand and simulate it we don't need to go below the level of pistons, cylinders, etc. Are there any cut-offs of this kind in the brain? The road map is not meant to offer answers, but I think after reading it one is inclined to think that there is probably a cut-off somewhere below neuronal level; you probably need to know about different kinds of neurotransmitters, but probably don't need to track individual molecules. Something like this seems to have been the level which the Blue Brain settled on.

The roadmap merely mentions some of the philosophical issues. It clearly has in mind the uploading of an individual consciousness into a computer, or the enhancement or extension of a biological brain by adding silicon chips, so an issue of some importance is whether personal identity could be preserved across this kind of change. If we made a computer copy of Stephen Hawking's brain at the moment of his death, would that be Stephen Hawking?

The usual problem in discussions of this issue is that it is easy to imagine two parallel scenarios; one in which Hawking dies at the moment of transition (perhaps the destruction of his brain is part of the process), and one in which the exact same simulation is created while he continues his normal life. In the first case, we might be inclined to think that the simulation was a continuation, in the latter case it's more difficult; yet the simulation in both cases is the same. My inclination is to think that the assertion of continuing identity in the first case is loose; we may choose to call it Hawking, but even if we do, we have to accept that it's Hawking put through a radical alteration.

Of course, even if the simulation hasn't got Hawking's personal identity, having a simulation of his brain (or even one which was only 80% faithful) would be a fairly awesome thing.

The roadmap provides a useful list of assumptions. One of these is:

Computability: brain activity is Turing computable, or if it is uncomputable, the uncomputable aspects have no functionally relevant effects on actual behaviour.

I've come to doubt that this is probable. I cannot present a rigorous case, but in sloppy impressionistic terms the problem is as follows. Non-computable problems like the halting problem or the tiling problem seem intuitively to involve processes which when tackled computationally go on forever without resolution. Human thought is able to deal with these issues by being able to 'see where things are going' without pursuing the process to the end.

Now it seems to me that the process of recognising meanings is very likely a matter of 'seeing where things are going' in much the same way. Computers don't deal with meaning at all, although there are cunning ploys to get round this in the various areas where it arises. The problem may well be that meanings are indefinitely ambiguous; there are always some more possible readings to be eliminated, and this might be why meaning is so untractable by computation.

Of course, apart from the hand-waving vagueness of that line of thought, it leaves me with the problem of explaining how the problem would manifest itself in the construction of a whole brain simulation; there would presumably have to be some properties of a biological brain which could never be accurately captured by a computational simulation. There are no doubt some fine details of the brain which could never be captured with perfect accuracy, but given the concept of scale separation, it's hard to see how that alone would be a fatal problem.

When a whole brain simulation is actually attempted, the answer will presumably emerge; alas, according to the estimates in the road map, I may not live to see it.

Is intentionality non-computable

22 November 2008

I undertook to flesh out my intuitive feeling that intentionality might in fact be a non-computable matter. It is a feeling rather than an argument, but let me set it out as clearly (and appealingly) as I can.

First, what do I mean by non-computability? I'm talking about the standard, uncontroversial variety of non-computability exhibited by the halting problem and various forms of tiling problem. The halting problem concerns computer programs. If we think about all possible programs, some run for a while and stop, while others run on forever (if for example there's a loop which causes the program to run through the same instructions over and over again indefinitely). The question is, is there any computational procedure which can tell which is which? Is there a program which, when given any other program, can tell whether it halts or not? The answer is no; it was famously proved by Turing that there is no such program, and that the problem is therefore undecidable or non-computable.

Some important qualifications should be mentioned. First, programs that stop can be identified computationally; you just have to run them and wait long enough. The problem arises with programs that don't halt; there is no general procedure by which we can identify them all. However, second, it's not the case that we can never identify a non-stopping program; some are obvious. Moreover, when we have identified a particular non-stopping program, we can write programs designed to spot that particular kind of non-stopping. I think this was the point Derek was making in his comment on the last point, when he asked for an example of a human solution that couldn't be simulated by computer; there is indeed no example of human discovery that couldn't be simulated - after the fact. But that's a bit like the blind man who claims he can find his way round a strange town just as well as anyone else; he just needs to be shown round first. We can actually come up with programs that are pretty good at spotting non-stopping programs for practical purposes; but never one that spots them all.

Tiling problems are really an alternative way of looking at the same issue. The problem here is, given a certain set of tiles, can we cover a flat plane with them indefinitely without any gaps? The original tiles used for this kind of problem were square with different colours, and an additional requirement was that colours must be matched where tiles met. At first glance, it looks as though different sets of tiles would fall into two groups; those that don't tile the plain at all, because the available combinations of colours can't be made to match up satisfactorily without gaps; and those that tile it with a repeating pattern. But this is not the case; in fact there are sets of tiles which will tile the plane, but only in such a way that the pattern never repeats. The early sets of tiles with this property were rather complex, but later Roger Penrose devised a non-square set which consists of only two tiles.

The existence of such 'non-periodic' tilings is the fly in the ointment which essentially makes it impossible to come up with a general algorithm for deciding whether or not a given set of tiles will tile the plane. Again, we can spot some that clearly don't, some that obviously do, and indeed some that demonstrably only do so non-periodically; but there is no general procedure that can deal with all cases.

I mentioned Roger Penrose; he, of course, has suggested that the mathematical insight or creativity which human beings use is provably a non-computable matter, and I believe it was the contemplation of how the human brain manages to spot whether a particular tricky set of tiles will tile the plane that led to this view (that's not to say that human brains have an infallible ability tell whether sets of tiles tile the plane, or computations halt). Penrose suggests that mathematical creativity arises in some way from quantum interactions in microtubules; others disagree with his theory entirely, arguing, for example, that the brain just has a very large set of different good algorithms which when deployed flexibly or in combination look like something non-computational.

I should like to take a slightly different tack. Let's consider the original frame problem. This was a problem for AI dealing with dynamic environments, where the position of objects, for example, might change. The program needed to keep track of things, so it needed to note when some factor had changed. It turned out, however, that it also needed to note all the things that hadn't changed, and the list of things to be noted at every moment could rapidly become unmanageable. Daniel Dennett, perhaps unintentionally, generalised this into a broader problem where a robot was paralysed by the combinatorial explosion of things to consider or to rule out at every step.

Aren't these problems in essence a matter of knowing when to stop, of being able to dismiss whole regions of possibility as irrelevant? Could we perhaps say the same of another notorious problem of cognitive science - Quine's famous problem of the impossibility of radical translation. We can never be sure what the word 'Gavagai' means, because the list of possible interpretations goes on forever. Yes, some of the interpretations are obviously absurd – but how do we know that? Isn't this, again, a question of somehow knowing when to stop, of being able to see that the process of considering whether 'Gavagai' means 'rabbit or more than two mice', 'rabbit or more than three mice' and so on isn't suddenly going to become interesting.

Quine's problem bears fairly directly on the problem of meaning, since the ability to see the meaning of a foreign word is not fundamentally different from the ability to see the meaning of words per se. And it seems to me a general property of intentionality, that to deal with it we have to know when to stop. When I point, the approximate line from my finger sweeps out an indefinitely large volume of space, and in principle anything in there could be what I mean; but we immediately pick out the salient object, beyond which we can tell the exploration isn't going anywhere worth visiting.

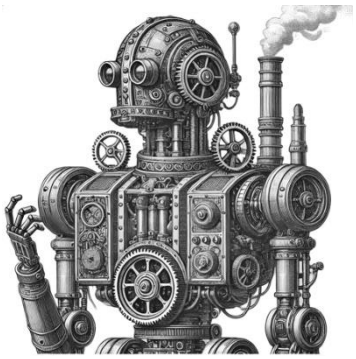
The suggestion I wanted to clarify, then, is that the same sort of ability to see where things are going underlies both our creative capacity to spot instances of programs that don't halt, or sets of tiles that cover the plane, and our ability to divine meanings and deal with intentionality. This would explain why computers have never been able to surmount their problems in this area and remain in essence as stolidly indifferent to real meaning as machines that never manipulated symbols.

Once again, I'm not suggesting that humans are infallible in dealing with meaning, nor that algorithms are useless. By providing scripts and sets of assumptions, we can improve the performance of AI in variable circumstances; by checking the other words in a piece of text, we can improve the ability of programs to do translation. But even if we could bring their performance up to a level where it superficially matched that of human beings, it seems there would still be radically different processes at work, processes that look non-computational.

Such is my feeling, at least; I certainly have no proof and no idea how the issue could even be formalised in a way that rendered it susceptible to proof. I suppose being difficult to formalise rather goes with the territory.

Ethical Kill-bots

7 December 2008



Robot ethics have been attracting media attention again recently. Could autonomous kill-bots be made to behave ethically - perhaps even more ethically than human beings?

Can robots be ethical agents at all? The obvious answer is no, because they aren't really agents; they don't make any genuine decisions, they just follow the instructions they have been given. They really are 'only following orders', and unlike human beings, they have no capacity to make judgements about whether to obey the rules or not, and no moral responsibility for what they do.

On the other hand, the robots in question are autonomous to a degree. The current examples, so far as I know, are relatively simple, but it's not impossible to imagine, at least, a robot which was bound by the interactions within its silicon only in the sort of way human beings are bound by the interactions within their neurons. After all it's still an open debate whether we ourselves, in the final analysis, make any decisions, or just act in obedience to the laws of biology and physics.

The autonomous kill-bots certainly raise qualms of a kind which seem possibly moral in nature. We may find land-mines morally repellent in some sense (and perhaps the abandonment of responsibility by the person who places them is part of that) but a robot which actually picks out its targets and aims a gun seems somehow worse (or does it?). I think part of the reason is the deeply embedded conviction that commission is worse than omission; that we are more to blame for killing someone then for failing to save their life. This feels morally right, but philosophically it's hard to argue for a clear distinction. Doctors apparently feel that injecting a patient in a persistent vegetative state with poison would be wrong, but that it's OK to fail to provide the nutrition which keeps the patient alive: rationally it's hard to explain what the material difference might be.

Suppose we had a more moral kind of land mine. It only goes off if heavy pressure is applied, so that it is unlikely to blow up wandering children, only heavy military vehicles. If anything, that seems better than the ordinary kind of mine; yet an automated machine gun which seeks out military targets on its own initiative seems somehow worse than a manual one. Rules which restrain seem good, while rules which allow the robot to kill people it could not have killed otherwise seem bad; unfortunately, it may be difficult to make the distinction. A kill-bot which picks out its own targets may be the result of giving new cybernetic powers to a gun which would otherwise sit idle, or it may be result of imposing some constraints on a bot which otherwise shoots everything that moves.

In practice the real challenge arises from the need to deal with messy reality. A kill-bot can easily be given a set of rules of engagement, required to run certain checks before firing, and made to observe appropriate restraint. It will follow the rules more rigorously than a human soldier, show no fear, and readily risk its own existence rather than breach the rules of engagement. In these respects, it may be true that the robot can exceed the human in propriety.

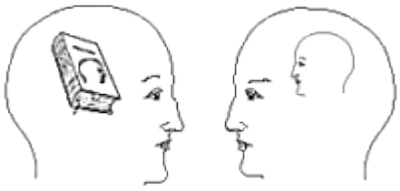
But practical morality comes in two parts; working out what principle to apply, and working out what the hell is going on. In the real world, even for human beings, the latter is the real problem more often than not. I may know perfectly clearly that it is my duty to go on defending a certain outpost until there ceases to be any military utility in doing so; but has that point been reached? Are the enemy so strong it is hopeless already? Will reinforcements arrive if I just hang on a bit longer? Have I done enough that my duty to myself now supersedes my duty to the unit? More fundamentally, is that the enemy, or a confused friendly unit, or partisans, or non-combatants? Are they trying to slip past, to retreat, to surrender, to annihilate me in particular? Am I here for a good reason, is my role important, or am I wasting my time, holding up an important manoeuvre, wasting ammunition? These questions are going to be much more difficult for the kill-bots to tackle.

Three things seem inevitable. There will be a growing number of people working on the highly interesting question of which algorithms produce, for any given set of computing and sensory limitations, the optimum ratio of dead enemies and saved innocents over a range of likely sets of circumstances. They will refer to the rules which emerge from their work as ethical, whether they really are or not. Finally, those algorithms will in turn condition our view of how human beings should behave in the same circumstances and affect our real moral perceptions. That doesn't sound too good, but again the issue is two-sided. Perhaps on some distant day the chief kill-bot, having absorbed and exhaustively considered the human commander's instructions for an opportunistic war will use the famous formula:

"I'm sorry Dave, I'm afraid I can't do that…"

Theory Theory and Other Theories

23 December 2008



Mitchell Herschbach spoke up for folk psychology in the JCS recently, suggesting that while there was truth in what its critics had said, it could not be dispensed with altogether.

What is folk psychology anyway? It is widely accepted that one of our basic mental faculties is the ability to understand other people - to attribute motives and feelings to them and change our own behaviour accordingly. One of the recognised stages of child development is the dawning recognition that other people may not know what we know, and may believe something different. There is relatively little evidence of this ability in other animals, although I believe chimps. for example, have been known to keep their discovery of food quiet if they thought other chimps couldn't see it. Commonly this ability to understand others is attributed to the possession of a 'theory of mind', or to the application of 'folk psychology', a set of commonsensical or intuitive rules about how people think and how it affects their likely behaviour.

So far as I'm aware, no-one has managed to set out exactly what 'folk psychology' amounts to. An interesting comparison would be the attempt some years ago to define folk physics, or 'naive physics' as it was called - it was thought that artificial intelligence might benefit from being taught to think the way human beings think about the real world, ie in mainly pre-Newtonian if not pre-Galilean terms. It proved easy enough to lay down a few of the laws of folk physics - bodies in motion tend to slow down and stop; heavy things fall faster than light ones - but the elaboration of the theory ran into the sand for two reasons; the folk-theory couldn't really be made to work as a deductive system, and as it developed it became increasingly complex and strange, until the claim that it resembled our intuitive beliefs became rather hard to accept. I imagine a comprehensive account of folk psychology might run into similar problems.

Of course, 'folk psychology' could take various forms besides a rigorously stated axiomatic deductive system.Herschbach explains that one of main divisions in the folk psychology camp is between the champions of theory theory (ie the idea that we really do use something relatively formal resembling a scientific theory) and simulation theory (you guessed it, the idea that we use a simulation of other people's thought processes instead). Some of course, are attracted by the idea that mirror neurons, those interesting cells that fire both when we perform action A and when we see action A performed by someone else, might have a role in providing a faculty of empathy which underpins our understanding of others. According to Herschbach the theory theorists and the simulation theorists have tended to draw together more recently, with most people accepting that the mind makes some use of both approaches.

However, the folk folk face a formidable enemy and a more fundamental attack. The 'phenomenological' critics of folk psychology think the whole enterprise is misguided; in order to guess what other people will do, we don't need to go through the rigmarole of examining their behaviour, consulting a theory and working out what their inward mental states are likely to be, then using the same theory to extrapolate what they are likely to do. We can deal with people quite competently without ever having to consider explicitly what their inner thoughts might be. Instead we can use 'online' intelligence, the sort of unreflecting, immediate understanding of life which governs most of our everyday behaviour.

The classical way of investigating the 'folk psychology' of children involves false-belief experiments. An experimenter places a toy in box a in full view of the child and an accomplice. The accomplice leaves the room and the experimenter moves the toy to box b. Then the child is asked where the accomplice will look for the toy. Younger children expect the accomplice to look for the toy where it actually is, in box b; when they get a little older they realise that the accomplice didn't see the transfer and therefore can be expected to have the false belief that the toy is still in box a. The child has demonstrated its ability to understand that other people have different beliefs which may affect their behaviour. Variations on this kind of test have been in use since Piaget at least.

Ha! say the critics, but what's going on here besides the main show? In addition to watching the accomplice and the toy, the child is going through complex interactions with the experimenter - obeying instructions, replying to questions and so on. Yet it would be absurd to maintain that they are carefully deducing the experimenter's state of mind at every stage. They just do as they're told. But if they can do all that without a theory of mind, why do they need one for the experiment? They just realise that people tend to look for things where they saw them last, without any nonsense about mental states.

Herschbach accepts that the case for 'online' understanding is good, but he thinks, briefly, that the opponents of folk psychology don't give sufficient attention to the real point of false-belief experiments. It may be that we don't have to retreat into self-conscious offline meditation to deal with false beliefs, but isn't it the case that online thinking is itself mentalistic to a degree?

It does seem unlikely that beliefs about other people's beliefs can be banished altogether from our account of how we deal with each other. In fact, our tendency to attribute beliefs to people by way of explanation is so strong we automatically do it even with inanimate objects which we know quite well have no beliefs at all ("MS Word has some weird ideas about grammar").

The problem, I think, is not that Herschbach is wrong, but that we seem to have ended up with a bit of a mess. In dealing with other people we may or may not use online or offline reasoning or some mixture of the two; our thinking may or may not be mentalistic in some degree, and it may or may not rely on a theory or a simulation or both. The brain is, of course, under no obligation to provide us with a simple, single module for dealing with people, but this tangle of possibilities seems so complicated we don't really seem to be left with any reliable insight at the end of it. Any mental faculty or way of thinking may, in an unpredictable host of different ways, be relevant to the way we understand each other. Alas (alas for philosophy of mind, anyway), I think that's probably the truth.

Radiotelepathy

15 January 2009

I see that at Edge they have kept up their annual custom of putting a carefully chosen question to a group of intellectuals. This year, they asked "What game-changing scientific ideas and developments do you expect to live to see?". There were many interesting answers. Freeman Dyson foresaw what he called 'radiotelepathy'. The idea is that a set of small implants record the activity of your brain which is then transmitted and delivered into someone else's (and vice versa). Hey presto, at once your thoughts and feelings are shared.

As the Thinking Meat Project remarked, this idea opens up a number of questions. Given our particular interests here, the first point that came to mind was that this kind of telepathy would surely resolve at last the vexed question of qualia. How do we know that the red seen by others looks the same as the red seen by us? Perhaps the experience they have when they see red is the experience we have when we see green? Or perhaps (a less-often discussed possibility) their red is a bit like our middle C on a badly tuned piano? Or perhaps their colour experiences are nothing like any of our experiences; perhaps there are an infinite number of phenomenal experiences which go with the perception of colour, and everyone has their own unique and ineffable set.

Well, with Dyson telepathy, there would be no need for us to wonder anymore; just tune in to someone else's brain, and we can have their experiences ourselves. Or can we? Perhaps not. Even as I write, I can sense hard-liners getting ready to insist that qualia recorded, transmitted, and inserted into a new brain, are not the same as the freshly gathered original ones. You still wouldn't know what the real thing was like. It might be that it is our brains themselves that impart the special what-it-is-likeness to experiences, in which case even telepathy won't help, and we can only ever have our own qualia. I think this exposes the insoluble problem at the heart of the whole qualia issue. Really the only way to know what someone's experiences are like in themselves, is to be that person. But you can't be someone else.

But steady on, because it seems highly unlikely to me that Dyson telepathy is feasible. I don't see any insoluble problem with the hardware he calls for, but downloading brain content is a tricky business, and uploading is even worse. To start with, Dyson talks about the 'entire brain', but do we want the whole thing? Do I want the activity of someone else's cerebellum reproduced in my own? Do I want the control routines for my caridovascular system overwritten? No thanks. So even on the macro scale we have to be very careful about where we put our ingoing signals. Pinpointing the right neurons seems a hopeless task. It's true that by and large the same regions of the cortex appear to deal with the same functions in different individuals, although variation is also quite possible. It's also true that recent research has identified individual neurons with very specific responses - neurons that fire, say, in response to the sight of Freeman Dyson, but not in response to anyone else. But so far as I know, it hasn't been demonstrated that the Dyson neuron in everyone is in the same place even approximately; it actually seems most unlikely, given that brains are wired in highly individual ways, and that indeed, most people have never had the pleasure of meeting Freeman Dyson. I don't think it's even been shown that the very same neuron which responds to Dyson today continues to do so

next week. Because all our machines are made to have their states encoded in a readable way, we tend to expect the same of nature, but evolution has no need of legible code. So it's very likely that the neural activity which in my brain corresponds to thinking of Freeman Dyson would, when transposed to another cranium, come out as the tantalising memory of the taste of a biscuit, or an intimation of mortality, or reservations about bicameralism.

Of course it's worse than that. Our brains are carefully organised, and the random dumping of alien activity would be oustandingly likely to mess things up. Would the brain activity that was there before - my own mental activity - be wiped out, so that instead of sharing someone else's thoughts, I suddenly thought I was someone else? Or could it somehow be merged? Dissolving the barriers and merging with another person can sound almost sensuously appealing (given the right person), but the sudden appearance of unforeshadowed alien thoughts might actually be terrifying, severely disorienting: a threat to the integrity of the psyche liable to end in trauma. In this respect, it's worth noting that nothing is more disruptive to one signal than another similar signal. If you write a sentence on a piece of paper and then cross it out with two or three lines, it remains easily legible. But if you write even one other sentence over the top of it, it becomes pretty much illegible at once. In the same way, it seems likely that activity from another brain would be the most dirsuptive thing you could input to your own, far worse than random noise.

I think the best alternative would be to home in on sensory inputs in the brain and try to place your interface somewhere early in the system before those inputs reach the more complex functions of consciousness. The result then would be more like hearing an external voice, or seeing an external hallucination. Much easier to deal with, but of course not so extraordinary - not really different in kind from using a video phone. At the end of the day, perhaps it's best to stick to the brain inputs provided by the designer - our normal senses. Walter Freeman memorably lamented the cognitive isolation in which we are all ultimately confined; but perhaps that isolation is the precondition of personal identity.

Fear in a handful of dust?

26 January 2009

Eric Schwitzgebel had an interesting post the other day asking why the universe isn't permeated with minds made of complex conjunctions of dust or other stuff. Most people would accept that 'brains' can in principle be made out of anything so long as the functional relationships are right; those relationships can be spread out over time and space, and so long as the right relationships are maintained, we don't even have to keep them in the right order. So shouldn't there be copies of your mind, and lots of other minds, spread all over the Universe? But that couldn't be true...

For a more careful exposition and an interesting discussion, see the original post. I think I fall on the sceptical side of this dialogue, more or less for some of the reasons articulated in the comments over there, so I won't repeat the points already made. Briefly, thought, consciousness, is surely a process, and if you split it up over time and space the process isn't really there. Never mind cogitation; would we even say that digestion could be instantiated by such a set-up? Here is a set of atoms; they are scattered over a huge area and thousands of years, but taken together in the right order they could constitute a steak and kidney pie; if you don't like that, perhaps the pie could appear momentarily, or even for several minutes, as the result of a bizarre but providential quantum accident inside my fridge, say. Then here's another set, or another bizarre occurrence, which duplicates the pie in a state of beginning to be bitten. And so on through to the unmentionable end. That's not really digestion, is it (and to be honest, perhaps not a totally accurate translation of Eric's argument either)?

But it's interesting to take a different tack and bite the bullet instead of the pie. Yes, OK, all that dust does have mental experience. But surely it only has the most tenuous and ethereal kind. To take another ridiculous example, suppose we were talking about an axe. Normally we require all the parts of an axe to be well co-ordinated in space and time, but we could have a dust-axe which was made up of different parts from different places and centuries. If we make the right selection of parts, we can have a series of axe-slices which even constitute its swinging and cutting wood. But it's not really a very good axe; its existence relies completely on the support of our imagination, even though it consists entirely of unobjectionable physical items. Its existence is thin and dependent; it's like the hrönir in the Borges story Tlön, Uqbar and Orbis Tertius, objects that exist only because someone thought they did. The experiences of the dust mind are almost as insubstantial as the experiences of fictional characters, or it seems as if it ought to be so.

But somehow, I find myself reluctant to say even that much; reluctant to grant the dust even the most evanescent form of experience. Why is that? It's because I feel a mind is something real, not just an abstraction which can arise as well out of a mistily conceptual object as out of a solid, fleshy brain. Strangely, I worry less about the axe because it's clear to me that whether something has axehood is a matter of how we use and think of it. Various misshapen stones can have a significant degree of axe-worthiness; you can say they are axes if that suits you and deny it another time. Surely you can't endow something with a mind and then take it away as your convenience requires?

But at the same time another part of my mind is taking the opposite tack. All this nonsense about pies and axes misses the point completely, because no-one supposes that such things are constituted by high-order functional relations. Minds, on the other hand, seem very likely to be entities of that kind, and so they are uniquely suited to be the abstract result of some reinterpretation of suitable sections of the world. It's just your hankering for some soul, some magic homunculus, that prevents you from realising this.

Come to that, what's the point of the dust? Suppose we couldn't find a suitable candidate for the role of one mote. OK, we say, it doesn't really matter, let's just arrange things so that the other parts of our dust mind behave as if this missing one were actually there. We can easily do that. But if we can remove one mote, why not remove them all? Let the functional relationships remain without physical realisation. If you're worried that imaginary dust has no causal powers, reflect that at this very moment it is having effects in making me describe it; even as I write, and you in some remote place and time read, it is rearranging in its insidious way the contents of both our minds. But if we can do without the physical realisation, space and time become irrelevant; the Universe is full of minds; in fact every point is a point of view, and Leibniz's Monadology is vindicated.

Once again I have reached that familliar state of complete confusion...

Are determinists evil?

1 February 2009

Normally we try to avoid casting aspersions on the character of those who hold a particular opinion; we like to take it for granted that everyone in the debate is honest, dispassionate, and blameless. But a recent paper by Baumeister, Masicampo and DeWall (2009), described in Psyblog, suggests that determinism (disbelief in free will) is associated with lower levels of helpfulness and higher levels of aggression. Another study reported in Cognitive Daily found that determinists are also cheats.

It's possible to question the way these experiments were done. They involved putting deterministic thoughts into some of the subjects' minds by, for example, reading them passages from the works of Francis Crick (who besides being an incorrigible opponent of free will in philosophical terms, also, I suppose, opened the way for genetic determinism). That's all very well, but it could be that, as it were, habitual determinists are better able to resist the morally corrosive effect of their beliefs than people who have recently been given a dose of persuasive determinism.

However, the results certainly chime with a well-established fear that our growing ability to explain human behaviour is tending to reduce our belief in responsibility, so that malefactors are able to escape punishment merely by quoting factors that influenced their behaviour. I was powerless; the crime was caused by chemical changes in my brain.

PsyBlog concludes that we must cling to belief in free will, which sounds perilously close to suggesting that we should pretend to believe in it even if we don't. But leaving aside for a moment the empirical question of whether determinists are morally worse than those who believe in free will, why should they be?

The problem arises because the traditional view of moral responsibility requires that the evil act must be freely chosen in order for the moral taint to rub off on the agent. If no act is ever freely chosen, we may do bad things but we shall never ourselves be truly bad, so moral rules have no particular force. A few determinists, perhaps, would bite this bullet and agree that morality is a delusion, but I think most would not. It would be possible for determinists to deny the requirement for freedom and say instead that people are guilty of wrongdoing simply when connected causally or in other specified ways with evil acts, regardless of whether their behaviour is free or not. This restores the validity of moral judgements and justifies punishment, although it leaves us curiously helpless. This tragic view was actually current in earlier times: Oedipus considered himself worthy of punishment even though he had had no knowledge of the crimes he was committing, and St Augustine had to argue against those who contended that the rape suffered by Lucretia made her a sinful adulteress - something which was evidently still a live issue in 1748 when Richardson was writing Clarissa, where the same point is raised. Even currently in legal theory we have the notion of strict liability, whereby people may be punished for things they had no control over (if you sell poisonous food, you're liable, even if it wasn't you

that caused it be poisonous). This is, I think a case of ancients and moderns reaching similar conclusions from almost antithetical understandings; in the ancient world you could be punished for things you couldn't have prevented because moral taint was so strong; in the contemporary world you can be punished for things you couldn't have prevented because moral taint is irrelevant, and punishment is merely a matter of deterrence.

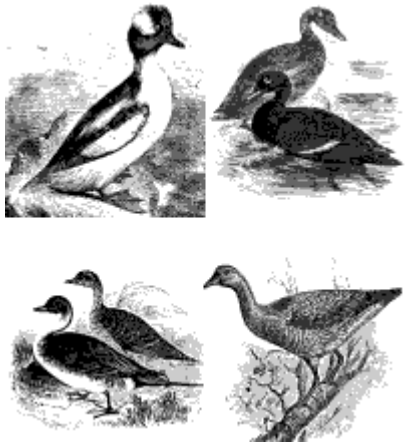
That is of course, the second escape route open to determinists; it's not about moral responsibility, it's about deterrence, social sanctions, and inbuilt behavioural norms, which together are enough to keep us all on the straight and narrow. This line of argument opens up an opportunity for the compatibilists, who can say: you evidently believe that human beings have some special capacity to change their behaviour in response to exhortation or punishment - why don't we just call that free will? More dangerously, it leaves the door open for the argument that those who believe their decisions have real moral consequence are likely to behave better than those who comply with social norms out of mere pragmatism and conditioning.

Meantime, to the rescue come De Brigard, Mandelbaum, and Ripley (pdf): as a matter of fact, they say, our experiments show that giving a neurological explanation for bad behaviour has no effect on people's inclination to condemn it. It seems to follow that determinism makes no difference. They are responding to Nahmias, who put forward the interesting idea of bypassing: people are granted moral immunity if they are thought to suffer from some condition that bypasses their normal decision-making apparatus, but not if they are subject to problems which are thought to leave that apparatus in charge. In particular, Nahmias found that subjects tended to dismiss psychological excuses, but accept neurological ones. De Brigard, Mandelbaum and Ripley, by contrast, found it made no difference to their subjects reactions whether a mental condition such as anosognosia was said to be psychological or neurological; the tendency to assign blame was much the same in both cases. I'm not sure their tests did enough to make sure the distinction between neurological and psychological explanations was understood by the subjects; but their research does underline a secondary implication of the other papers; that most people are not consistent and can adopt different interpretations on different occasions (notably there were signs that subjects were more inclined to assign blame where the offence was more unpleasant, which is illogical but perhaps intuitively understandable).

I suspect that people's real-life moral judgements are for the most part not much affected by the view they take on a philosophical level, and that modern scientific determinism has really only provided a new vocabulary for defence lawyers. A hundred or two hundred years ago, they might have reminded a jury of the powerful effect of Satan's wiles on an innocent but redeemable mind; now it may be the correctable impact of a surge of dopamine they prefer to put forward.

New Turing Tests

9 February 2009



The New Scientist, under the title ‘Tests that show machines closing in on human abilities’ has a short review of some different ways in which robotic or computer performance is being tested against human achievement. The piece is not really reporting fresh progress in the way its title suggests - the success of Weizenbaum’s early chat-bot Eliza is not exactly breaking news, for example - but I think the overall point is a good one. In the last half of the last century, it was often assumed that progress towards AI would see all the barriers come down at once: as soon as we had a robot that could meet Turing’s target of a few minutes’ successful small-talk, fully-functioning androids would rapidly follow.

In practice, Turing’s test has not been passed as he expected, although some stalwart souls continue to work on it. But we have seen the overall problem of human mentality unpicked into a series of substantial, but lesser challenges. Human levels of competence remain the target, but competence in different narrow fields, with no expectation that solving the problem in one area solves it in all of them.

The piece ends with a quote from Stevan Harnad which suggests he clings to the old way of looking at this:

“If a machine can prove indistinguishable from a human, we should award it the respect we would to a human”

That may be true, in fact, but the question is, indistinguishable in which respects? People often quote a particular saying in this respect: if it walks like a duck and quacks like a duck, it’s a duck. Actually, even in the case of ducks this isn’t as straightforward as it might seem since, other wildfowl may be ducklike in some respects. Given a particular bird, how many of us could say with any certainty whether it were a Velvet Scoter, a White-winged Coot - or just a large sea-duck? But it’s worse than that. The Duck Law, as we may call it, works fairly well in real life; but that’s because as a matter of fact there aren’t all that many anatine entities in the world other than outright ducks. If there were a cunning artificer who was bent on turning out false, mechanical ducks like the legendary one made by Vaucanson, which did not merely walk and quack like a duck, but ate and pooped like one, we should need a more searching principle. When it comes to the Turing Test, that is pretty much the position we find ourselves in.

There is, of course, a more rigorous version of Duck Law which is intellectually irreproachable, namely Leibniz’s Law. Loosely, this says that if two object share the same properties, they are the same. The problem is, that in order to work, Leibniz’s law has to be applied in the most rigorous fashion. It requires that all properties must be the same. To be indistinguishable from a human being in this sense means literally indistinguishable, ie having human guts, a human mother and so on.

So, in which respects must a robot resemble a human being in order to be awarded the same respect as a human? It now seems unlikely that a machine will pass the original Turing Test soon; but even if it did, would that really be enough? Just looking like a human, even in flexible

animation which reproduces the pores of the skin and typical human movements, is clearly not enough. Nor is being able to improvise jazz tunes seamlessly with a human player. But these things are all significant achievements nevertheless. Perhaps this is the way we make progress.

Or possibly, at some stage in the future, someone will notice that if he were to bolt together a hundred different software modules and some appropriate bits of hardware, all by then readily available, he could theoretically produce a machine able to do everything a human can do; but he won't by then see any point in actually doing it.

FORs, HORs, and unicorns

28 February 2009

Pete Mandik has an intriguing argument (expounded in the recent JCS) which he has named after the unicorn. It's an argument against certain kinds of theory of consciousness, and its surprising central premise is that there is no such thing as the property of being represented.

How can that be? Properties come for free, don't they - in my arguments at least, I can have any property I can think of for the asking surely? And if there is no such thing as the property of being represented, it seems to follow inevitably that nothing is being represented, nothing ever has been represented, and nothing ever will be represented. That is a trifle counter-intuitive, to say the least.

The centre of the argument is the contention that since there are mental representations of things that don't exist (such as unicorns), there can't be any such property as being represented. Curiouser and curiouser: what's the problem with non-existent things having properties? Don't unicorns have the properties of being equine, and horned, and for that matter, mythical? If they don't, we seem to have some problems on our hands. How are we going to tell the difference between unicorns and hippogriffs, which on this view have exactly the same properties (ie none)? Although I suppose it must be granted that telling the difference between a cage with all the hippogriffs in the world in it, and one containing all the unicorns, might be tricky.

We need, of course, to see the argument in context, since it is really directed narrowly at two specific kinds of theory. The first kind are Higher Order Representation (HOR) theories. These say, in essence, that a conscious mental state is one we are in turn conscious of - or to avoid the infinite regress which seems to threaten there, a conscious thought is one we're aware of having, we might say. The second kind, First Order Representational (FOR) theories say that for us to be conscious of something, for it to be phenomenal, it has to be represented appropriately in our awareness. Both of these kinds of theory must fall, says Mandik, if we pull away the carpet of representedness from under them.

By way of background, Mandik sets out rather nicely what he calls 'The problem of Intentionality'. It amounts to the incompatibility of three plausible propositions:

1. Relations can only obtain between relata that exist.
2. There exist mental representations of nonexistent things.
3. Representation is a relation between that which does the representing and that which is represented.

I suggested above that it might be natural to question proposition 1; but Mandik prefers to give up 3. As he makes clear in his conclusion, he's OK with representing, it's just the being represented which is the problem. I think the intuition behind this is now clearer; representation is something that appertains to the representer; it doesn't really touch the represented.

This makes some sense. One of the puzzling features of intentionality is its seemingly unlimited power to reach across time and space in a most peculiar way. It's easy for me to pick out

someone distant in time and space and refer to them. Indeed, it can be someone of whom I know virtually nothing. Take the 10,000th man who ever saw a picture of a unicorn. He surely existed, but neither I nor he could identify him. Nevertheless, I referred to him successfully; I can say if I like that the extra full stop at the end of this sentence represents him. There he is. I've suddenly changed his properties, since he now has the additional property of being one of the select company of people referred to in this blog, though sadly he never knew it.

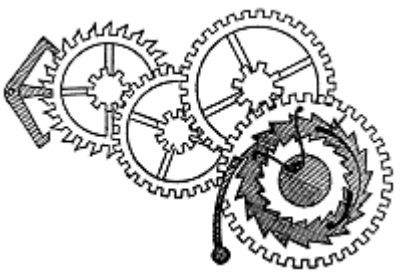
Nonsense! Or so I suppose Mandik would say; you haven't touched that man at all. The property of 'being mentioned here' is not really one of his properties in any meaningful sense, and there are no real relations between him and you. His being represented in this sort of sense just is not the kind of real property which could provide the basis for your being conscious of him.

If you doubt it, he might add, reprising a point made in the paper, what about square circles? Or, say, all those people not referred to here (paradoxical because that phrase itself refers to them). We can talk about them if we like, but surely it's clear that talking about them is not to stand in any real relation to them. And if it doesn't work for them, it doesn't work for anything or anyone.

This is an intriguing argument, and although the current paper only concerns itself with HORs and FORs, it obviously has implications in a much wider context. There are of course, other arguments which can also be deployed against FORs and HORs. In passing, Mandik offers a nicely deflating explanation of the appeal of higher-order theories. One thing that's true about thoughts we're not aware of, he points out, is that we're not aware of them. Consequently, when we introspect, it's not surprising if all our conscious states seem to be ones we're conscious of

On the Level about Computers and Minds

15 March 2009



Ari N. Schulman had an interesting piece in the New Atlantis recently on *Why Minds Are Not Like Computers*. Very briefly his view is that the aims of the strong AI project have quietly become less ambitious over time. In particular, from aiming to find the algorithms which directly generate high-level consciousness in one fell swoop, researchers have turned to simulating lower-level mechanisms and modules; in some cases they've gone to a further level and are attempting to simulate the brain at neuronal level. Who knows, they might end up trying to do it at molecular or subatomic level, he says, but the point is that at these low levels the game is already lost; even if the simulation runs, we still won't understand what's happening on the higher levels. If we have to go to the level of simulating individual neurons, the original claim that the brain works the way a computer does has implicitly been abandoned.

Schulman thinks that a misleading application of an analogy between computers and the mind is key to the problem. In computers we have the well-established set of layers which goes from high-level languages through machine code all the way down to physical transistors; researchers assumed, he thinks, that they would in effect be able to reverse engineer the source code of the brain and come up with high-level scripts which explicated the mechanisms of consciousness; and that they would be able to do it simply by ensuring that the input-output relationships were reproduced without having to worry about whether the hidden inner mechanisms of their version were actually the same as those of nature. But it never happened, and as time goes by it seems less and less likely that those algorithms will ever be found; it looks as if the mind just isn't like that after all.

Although there's something recognisable about that, I'm not sure whether this is a completely accurate account historically. It is certainly true that the misplaced early optimism of Good Old Fashioned Artificial Intelligence is a thing of the past - but that's hardly breaking news. I'm not sure that even in the most upbeat period people thought that they could do the entire mind in one go: even then they surely looked to start with simplified tasks and build single modules. It's just that they thought these modules and tasks would be dealt with more quickly and easily than they really were. The recent emergence of projects like the Blue Brain, seeking to simulate the neuronal level working of the brain, are also less a sign of lack of confidence in AI and more a sign of growing confidence in the number-crunching power of the computers now available. I don't think such projects are exactly typical of where things are at these days in any case.

Still, plausible claims? Of course, working out the high-level code, or even recognising the general drift of a program, from looking only at the machine-code level is not at all an easy business, so the fact that looking at neurons for many years has not yet led us to a general theory of consciousness is not necessarily a sign that it never will. Schulman does not, like Searle (whose views he discusses), take the view that something about the physical stuff of brains is essential; his objection to functionalism seems merely to be that it hasn't worked yet. Perhaps we just need more patience.

We also need to be a little careful about the diagnosis. There are actually different ways of dividing the whole business into levels; one is the programmer-facing way which Schulman

mainly focuses on; another is the user-facing one. Here the bottom level is made up of meaningless symbols; somewhere in the middle is organised data; and at the top is meaningful information. Surely it's here that the aim of AI researchers has been focussed; they expected consciousness, not as high-level program code, but as outputs which mean something to a human being, or which 'make sense' in the context of a task. Ultimately, for consciousness, the outputs have to make sense to the machine itself. If some form of computationalism can deliver these results, I don't think the absence of a high-level theory in the other sense would indicate philosophical failure.

Even if we do ultimately have to go beyond a narrow functionalist view, we need not abandon the overall quest. We should perhaps hang on to the distinction between consciousness as computation and consciousness as computable. The idea that the mind actually is just the programs running in the brain may look less plausible than it did; the idea that programs running somewhere might sustain a mind might yet be getting a second wind. It might be too narrow a view of clocks to say that they are nothing more than cogs, springs, and other pieces of mechanism; the works don't tell us what the essence of a clock is. But the ironmongery is all we need to make a clock, and perhaps we could make one that worked before we fully understood the principle of the escapement mechanism…?

Cryptic Consciousness

21 March 2009

I was thinking about the New Mysterian position the other day, and it occurred to me that there are some scary implications which I, at any rate, had never noticed before.

As you may know, the New Mysterian position, cogently set out by Colin McGinn, is that our minds may simply not be equipped to understand consciousness. Not because it is in itself magic or inexplicable, but because our brains just don't work in the necessary way. We suffer from cognitive closure. Closure here means that we have a limited repertoire of mental operations; using them in sequence or combination will take us to all sorts of conceptual places, but only within a certain closed domain. Outside that domain there are perfectly valid and straightforward ideas which we can simply never reach, and unfortunately one or more of these unreachable ideas is required in order to understand consciousness.

I don't think that is actually the case, but the possibility is undeniable; I must admit that personally there's a certain element of optimistic faith involved in my rejection of Mysterianism. I just don't want to give up on the possibility of a satisfying answer.

Anyway, suppose we do suffer from some form of cognitive closure (it's certainly true that human minds have their limitations, notably in the complexity of the ideas they can entertain at any one time). One implication is that we can conceive of a being whose repertoire of mental operations might be different from ours. It could be a god-like being which understands everything we understand, and other things besides; its mental domain might be an overlapping territory, not very different in extent from ours; or it might deal in a set of ideas all of which are inaccessible to us, and find ours equally unthinkable.

That conclusion in itself is no more than a somewhat frustrating footnote to Mysterianism; but there's worse to follow. It seems just about inevitable to me that if we encountered a being with the last-mentioned fully cryptic kind of consciousness, we would not recognise it. We wouldn't realise it was conscious: we probably wouldn't recognise that it was an agent of any kind. We recognise consciousness and intelligence in others because we can infer the drift of their thoughts from their speech and behaviour and recognise their cogency. In the case of the cryptics, their behaviour would be so incomprehensible, we wouldn't even recognise it as behaviour.

So it could be that now and then in AI labs a spark of cryptic consciousness has flashed and died without ever being noticed. This is not quite as bonkers as it sounds. It is apparently the case that when computers were used to apply a brute force, exhaustive search was applied to a number of end-game positions in chess, it turned out that several which had long been accepted as draws turned out to have winning strategies (if anyone can provide more details of this research I'd be grateful – my only source is the Oxford Companion to the Mind) . The strategies proved incomprehensible; to human eyes, even those of expert chess players; the computer appeared merely to bimble about with its pieces in a purposeless way, until unexpectedly, after a prolonged series of moves, checkmate emerged. Let's suppose a chess-playing program had independently come up with this strategy and begun playing it. Long before

checkmate emerged – or perhaps even afterwards – the human in charge would have lost patience with the endless bimbling (in a position known to be a draw, after all), and withdrawn the program for retraining or recoding.

Perhaps, for that matter, there are cryptically conscious entities on other planets elsewhere in the Galaxy. The idea that aliens might be incomprehensible is not new, but here there is a reasonable counter-argument. All forms of life, presumably, are going to be the product of a struggle for survival similar to the one which produced us on Earth. Any entity which has come through millions of years of such a struggle is going to have to have acquired certain key cognitive abilities, and these at least will surely be held in common. Certain basic categories of thought and communication are surely going to be recognisable; threats, invitations, requests, and the like are surely indispensable to the conduct of any reasonably complex life, and so even if there are differences at the margins, there will be a basis for communication. We may or may not have a set of cognitive tools which address a closed domain – but we certainly haven't got a random selection of cognitive tools.

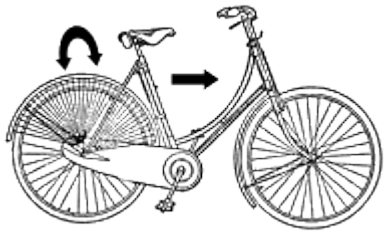
That's a convincing argument, though not totally conclusive. On Earth, the basic body plans of most animal groups were determined long ago; a few good basic designs triumphed and most animals alive today are variations on one of these themes. But it's possible that if things had been different we might have emerged with a somewhat different set of basic blueprints. Perhaps there are completely different designs on other planets; perhaps there are phyla full of animals with wheels, say. Hard to be sure how likely that is, because we only have one planet to go on. But at any rate, if that much is true of body plans, the same is likely to be true of Earthbound minds; a few basic mental architectures that seemed to work got established way back in history, and everything since is a variation. But perhaps radically different mental set-ups would have worked equally well in ways we can't even imagine, and perhaps on other worlds, they do.

The same negative argument doesn't apply to artificial intelligence, of course, since AI generally does not have to be the result of thousands of generations of evolution and can jump to positions which can't be reached by any coherent evolutionary path.

Common sense tells us that the whole idea of cryptic consciousness is more of a speculative possibility than a serious hypothesis about reality – but I see no easy way to rule it out. Never mind the AI labs and the alien planets; it's possible in principle that the animists are right and that we're surrounded every day by conscious entities we simply don't recognise…

Abilities Don't Matter

14 April 2009



Sam Coleman returned in a recent paper (JCS vol 16, n0 2-3) to the old perennial of Mary the colour scientist. As you may know, Frank Jackson's story about Mary was intended to establish that there was something important - qualia - which the simple physical account of the world omitted. Mary, we're told, knows everything that could possibly be known about colour from a scientific point of view (far more than any living scientist knows, or could know). She knows all the physical facts. But she has never experienced colour; when she sees a red rose for the first time, doesn't she learn something new - ie what red looks like? If you agree, then, it is argued, you must agree that there are things in heaven and earth not dreamt of in physicalist philosophy.

One standard riposte to this line of argument is to deny that Mary learns any new facts when she sees red for the first time: instead, she merely acquires a new ability. Now she can recognise red, when she couldn't do before; but she doesn't actually know any new facts about redness or the human optical system. (After all, the phrase 'she knows what red is like' literally suggests that she can give a list of things which are like a red object, and say whether the colour of a given object resembles red or not - which sounds a lot like recognition.) We might mention other abilities; the ability to remember or imagine redness, in particular - but none of these involve new knowledge in the ordinary sense, any more than acquiring the ability to ride a bike involves learning new facts. You can read as many books about bicycles as you like: you'll still fall off the first time you get on one.

Coleman aims to knock this line of argument on the head, not by refuting it but by showing it to be irrelevant. Most attacks on the Ability Hypothesis, he says, address what he calls its inner face directly. They seek to show that what happens to Mary cannot be boiled down to the gaining of an ability - perhaps because all such abilities involve factual knowledge - or they set out to show directly that Mary does indeed learn new facts. These strategies tend, he says with some justice, to end up mired in a clash of intuitions with no clear way forward.

Instead, then, he concedes that if we wish, we can see what happens to Mary as the acquisition of some new abilities; he addresses instead what he calls the outer face of the problem. What about the analogy of bike-riding? Proponents of the Ability Hypothesis say it does not depend on factual knowledge, and it's true it doesn't come from academic book-reading. But what does it involve? Surely it's all about keeping our balance? That's a matter of teaching your muscles how to respond quickly and appropriately to certain sensations of tipping over. So hang on, it actually involves knowledge of what it feels like to wobble - what certain kinds of phenomenal experience are. Is this kind of knowledge factual knowledge? Never mind the answer for the moment - all we need do is notice that this is exactly the kind of question we were asking about Mary's experience of redness in the first place. The Ability Hypothesis has merely taken us back to where we started - so we need not waste our time on it.

A neat piece of footwork, I think, and Coleman's analysis of the phenomenal element in abilities such as bike-riding, ear-waggling, and the use of chopsticks provides an interesting new insight, worth having in its own right. (Readers may be interested in the Phenomenal Qualities Project,

whose three-year mission is due to begin next month; Coleman is co-investigator to Paul Coates and the list of project members includes some stellar names.)

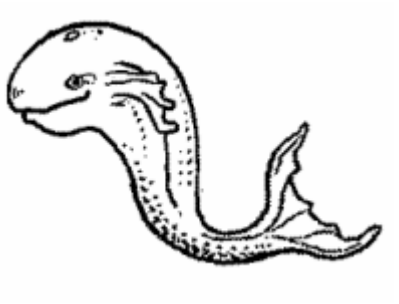
Where are we left with Mary meanwhile, though? Should we now accept the refutation of physicalism she represents? No, I would say, but then again yes.

One thing that's sure about Mary is that if she does acquire new knowledge, it is knowledge of some special kind. It isn't the sort we can write down and transmit in words - that much is true ex hypothesi, because if it were that kind of knowledge, it would be among the things Mary already knows. It's tempting to denounce this kind of knowledge as metaphorical or worse, but let's merely ask whether it's fair to expect the theory of physicalism to include it. Theories, after all, are expressed in words, and typically written down; if we're dealing with facts that can't be dealt with that way it seems only reasonable to conclude that no theory could deal with them and that they must remain extra-theoretical. Physicalism might then be incomplete in some sense, but we have no reason to think it isn't as complete as any theory could be. How much is it reasonable to ask?

And yet, and yet… We're talking here about what things are like; this somehow still feels like a problem where more could be said after all. We can't expect that our theories should deliver ineffable reality itself, but doesn't it seem that there is some tantalising last elucidation obtainable here, if only we knew how to go about it?

Forget AI

3 May 2009



In recent years there seems to have been a general trend in AI research towards more narrow and perhaps more realistic sets of goals; towards achieving particular skills and designing particular modules tied to specific tasks rather than confronting the grand problem of consciousness itself. The proponents of AGI feel that this has gone so far that the terms 'artificial intelligence' and AI no longer really designate the topic they're interested in, the topic of real thinking machines. 'An AI' these days is more likely to refer to the bits of code which direct the hostile goons in a first-person shooter game than to anything with aspirations to real awareness, or even real intelligence.

The mention of 'real intelligence' of course, reminds us that plenty of other terms have been knocked out of shape over the years in this field. It is an old complaint from AI sceptics that roboteers keep grabbing items of psychological vocabulary and redefining them as something simpler and more computable. The claim that machines can learn, for example, remains controversial to some, who would insist that real learning involves understanding, while others don't see how else you would describe the behaviour of a machine that gathers data and modifies its own behaviour as a result.

I think there is a kind of continuum here, from claims it seems hard to reject to those it seems bonkers to accept rather like this...

Claim: machines...	**Objection**
add numbers.	Really the 'numbers' are a human interpretation of meaningless switching operations.
control factory machines.	Control implies foresight and intentions whereas machines just follow a set of instructions.
play chess.	Playing a game involves expectations and social interaction, which machines don't really have.
hold conversations	Chat-bots merely reshuffle set phrases to give the impression of understanding.
react emotionally	There may be machines that display smiley faces or even operate in different 'emotional' modes, but none of that touches the real business of emotions.

Readers will probably find it easy to improve on this list, but you get the gist. Although there's something in even the first objection, it seems pointless to me to deny that machines can do addition - and equally pointless to claim that any existing machine experiences emotions - although I don't rule even that idea out of consideration forever.

I think the most natural reaction is to conclude that in all such cases, but especially in the middling ones, there are two different senses - there's playing chess and really playing chess. What annoys the sceptics is their perception that AIers have often stolen terms for the easy computable sense when the normal reading is the difficult one laden with understanding, intentionality and affect.

But is this phenomenon not simply an example of the redefinition of terms which science has always introduced? We no longer call whales fish, because biologists decided it made sense to make fish and mammals exclusive categories - although people had been calling whales fish on and off for a long time before that. Aren't the sceptics on this like diehard whalefishers? Hey, they say, you claimed to be elucidating the nature of fish, but all you've done is make it easy for yourself by making the word apply just to piscine fish, the easy ones to deal with. The difficult problem of elucidating the deeper fishiness remains untouched!

The analogy is debatable, but it could be claimed that redefinitions of 'intelligence' and 'learning' have actually helped to clarify important distinctions in broadly the way that excluding the whales helped with biological taxonomy. However, I think it's hard to deny that there has also at times been a certain dilution going on. This kind of thing is not unique to consciousness - look what happened to 'virtual reality', which started out as quite a demanding concept, and was soon being used as a marketing term for any program with slight pretensions to 3D graphics.

Anyway, given all that background it would be understandable if the sceptical camp took some pleasure in the idea that the AI people have finally been hoist with their own petard, and that just as the sceptics, over the years, have been forced to talk about 'real intelligence' and 'human-level awareness', the robot builders now have to talk about 'artificial general intelligence'.

But you can't help warming to people who want to take on the big challenge. It was the bold advent of the original AI project which really brought consciousness back on to the agenda of all the other disciplines, and the challenge of computer thought which injected a new burst of creative energy into the philosophy of mind, to take just one example. I think even the sceptics might tacitly feel that things would be a little quiet without the 'rude mechanicals': if AGI means they're back and spoiling for a fight, who could forbear to cheer?

'Supersized Mind'

17 May 2009

I was agreeably surprised by Andy Clark's 'Supersizing the Mind'. I had assumed it would be a fuller treatment of the themes set out in 'The Extended Mind', the paper he wrote with David Chalmers, and which is included in the book as an Appendix. In fact, it ranges more widely and has a number of interesting points to make on the general significance of embodiment and mind extension. Various flavours of externalism, the doctrine that the mind ain't in the head, seem to be popular at the moment, but Clark's philosophical views are clearly just part of a coherent general outlook on cognition.

Early on, Clark contrasts different approaches to the problem of robotic walking. Asimo, Honda's famous robot, does a pretty impressive job of walking, but achieves it through an awful lot of very carefully controlled mechanical kit. The humanoid shape of the body, apart from giving the robot its anthropomorphic appeal, is treated as more or less an incidental feature. By contrast, Clark cites Collins, Wisse and Ruina on passive dynamic walkers whose design gives them a natural tendency to waddle along with no control system at all. These machines are far less complex (one walking toy patented by Fallis in 1888 consists entirely of two pieces of cleverly bent wire - I'm going to try making my own later), but they embody a different kind of sophistication. As always, evolution got there first, and the relatively good levels of energy efficiency achieved by the ambulant human body, in contrast to the energy-guzzling needs of an Asimo-style approach, show that although human walking is more controlled than two pieces of wire, walking has been cunningly built into the design.

All this serves to introduce the idea that form has more importance than we may think, that legs are not just tools strapped on to the bottom of a standard unit. The example of feeling with a stick is often quoted; when we use a stick to probe the texture and shape of an object, it doesn't feel as if we're registering the motion of the stick against our hand, but rather as if we were really feeling with the stick, as though the stick itself had for the moment become a sensory extension, with feeling at the end, not in the handle.

Clark wants to say that this feeling is not an illusion, that the stick, and other extensions we may use, really are incorporated into us temporarily. He quotes a dystopian vision of Bruce Sterling in which powerful machines are driven by increasingly feeble and senile humans; why instead shouldn't the combination of feeble human and carefully designed robotics amount instead to a newly invigorated and intelligent person?

Among the most powerful extensions available for human beings is, of course, language. Words are often considered solely as a means of communication, but he makes a convincing case for the conclusion that they actually add massively to our cognitive abilities. He rightly says that the complexity generated by our ability to think about the way we think, or to think of new worlds from which other new worlds can be found, is staggering.

One of the ways our extended cognition helps us to extend ourselves further is by the construction of niches - re-engineering parts of the world to enhance our own performance. Clark here gives a fascinating account of how Elizabethan actors coped with the impossible load imposed on their memories - in those days it was apparently customary for a company to

play six different plays a week with few repetitions and entirely new plays arriving at regular intervals.

Key factors were the standard layout of the stage and documents known as ‘plots’, a kind of high-level map of who comes in when and where and does what. These would enable actors to get a quick overall grasp of a new play, and with a script which gave only their own lines and using an over-learned grasp of the dramatic conventions of the day which gave them an in-built knowledge of the sort of thing that could be expected, they were be able to pick up plays and switch from one to another almost on the fly as they went. This is just a particularly striking example of the way we often make things easier for our brains by reducing the epistemic complexity of the environment. Putting that piece of IKEA furniture together may be tough, but if we take out all the components and lay them out neatly the way they’re shown in the diagram, it actually becomes easier.

Clark draws a useful distinction, in many ways a key point, between three grades of embodiment. In mere embodiment, the body and sense are simply pieces of equipment with which things can be done; in basic embodiment their own feature and dynamics become resources in themselves which the entity in question can exploit. Finally, in the profound embodiment which humans and some other animals, especially primates, have developed, new parts of the world and new uses of parts of the body are constantly discovered and recruited to serve previously unimagined purposes. This capacity to pick up things and make them part of your own mental apparatus might be a defining feature of human cognition and human consciousness.

All this is fine, but we may ask whether it is enough to make us think cognitive extension is anything more than some fancy metaphors about plain old tool use. Clark now goes on to tackle some objections, first those made by Adams and Aizawa, who in a nutshell say that the kind of processes Clark are talking about are just not the right kind of thing to be considered cognitive. Various grounds are advanced, but the strongest part of the negative case is perhaps the claim that cognitive processes are distinguished by processes involving non-derived representations. In other words, real minds deal in stuff which is inherently meaningful, and the kinds of things Clark is on about, if they deal in meaning at all, do so only because a real brain interprets them them as doing so. This claim clearly rebuts the example of Otto, used by Clark and Chalmers: Otto, briefly, uses a notebook as a supplementary memory. It’s not memory, or memory-like, Adams and Aizawa would say, because the information in Otto’s brain is live, it means something all on its own, whereas his notebook only means something through the convention of writing and when brains write or read in it.

Clark suggests that meaningfulness, non-derived representations, can perhaps be found independently of brains in some cases. This is highly contentious territory where I think it might actually be more prudent not to go. There’s no need in any case, since he also has the argument that brains frequently use arbitrary, conventional symbols in internal cognitive activities - whenever we use words or numbers to think with. If those processes are cognitive, surely similar ones using external symbols deserve the same recognition. Is there any real difference between someone who visualises a map in memory in order to work out where to go, and someone who looks at the paper equivalent, such that we can dismiss one as not cognitive?

A different challenge comes from Robert Rupert, who proposes that embedded cognition is enough: we can say that cognitive processes are heavily dependent on external props and prompts without having to make the commitments required by extended cognition. The props

and prompts may be essential, but that doesn't mean we have to bring them within the pale of cognition itself - and in fact if we do we may risk losing sight of the thing we set out to study in the first place. If the mind includes all the teeming series of objects we use to help us think clearly, then the science of the mind is going to be more intractable than ever.

This latter objection is easily dealt with, since Clark is not concerned to deny that there is a recognisable, continuing core of cognition in the brain. On the other point, I'm not sure that choosing embedding over extension has any real advantage here: embedding might give us a cleanly delineated target for the study of the mind, but it explicitly does so by narrowing the scope of that study to exclude all the allegedly intractable swarm of prompts and props which help us out. But aren't the props and prompts an interesting and essential area of study, even if we deny them the status of being truly cognitive? Clark seems to accept that there is no knock-out case for extension rather than embedding, but he claims a number of advantages for the former, notably that it steers us safely away from 'magic dust' and the seductive old idea of the inner homunculus, the little man in our head who controls everything.

In a final section, Clark considers two 'Limits of Embodiment'. The first has to do with Noë's argument that perception is active; that seeing is like painting. This Strong Sensorimotor model has many virtues in Clark's eyes: in particular it moves attention away from the vexed issue of qualia and on to skills. In fact, it is associated with the claim that experience is just constituted by the activity of perception, which leaves no ambiguous ineffable inner reality to trouble us, dismissing the menacing hordes of philosophical zombies that qualophiles would unleash on us. The snag, in Clark's eyes, is 'hypersensitivity' - if we take this line at face value we find ourselves saying that only identical people could have identical experiences, and therefore that no actual people see the same things.

The other limit has to do with an argument that if embodiment is true, functionalism must be false. It is a basic claim of functionalism, after all, that the substrate doesn't matter; that in principle you could make a brain out of anything so long as it had the right functional properties at a high level. But embodiment says that the nature of the substrate is crucial - so if one is true the other must be false, right? Put as baldly as that, I think the weakness of the argument is apparent; why shouldn't it be the case that the things which the theory of embodiment says are crucial are in fact, in the final analysis, functional properties?

The book is a fascinating read, and full of stuff which this short sketch can't really do any justice to. I don't know whether I've become a convert to mind extension, though. The problem for me at a philosophical level, is that none of this really worries me. There are a number of issues to do with consciousness and the mind where my inability to see the answers is, in a quiet but deep way, quite troubling to me. Cognitive extension, somehow not so much.

I suppose part of the reason is that lurking at the back of my mind is the feeling that the location of the mind is a false issue. Where is the mind? In some Platonic realm, or zone of abstraction, or perhaps the question is meaningless. When we ask, is the mind within the brain or in the external world , the answer is in fact 'No'. Of course I understand that Clark is not operating at this naive level, but it leaves a lingering doubt somewhere at the back of my mind about how much the answer really matters , which is reinforced by the apparent softness of the issue. Clark himself seems to concede that it's not exactly a question of absolute right and wrong, more of whether this is a useful and enlightening way to look at things. On that level, I certainly found his account convincing.

Perhaps I’ve ended up becoming one of those frivolous people who always want a theory to boggle their minds a bit. It seems in the final analysis less a matter of truth than whether this a useful way of looking at the issue. Clark. to me, is certainly convincing on that.

Panpsychist Consciousness

31 May 2009

The JCS has devoted its latest issue to definitions of consciousness. I thought I'd done reasonably well by quoting seventeen different views, but Ram L. P. Vimal lists forty, in what he acknowledges is not a comprehensive list. There is much to be said about all this - and Bill Faw promises a book-length treatment of the thoughts offered in his paper - but much of the ground has been trodden before.

A notable exception is David Skrbina's panpsychist view. I have been accused in the past of being unfair to panpsychism, the belief that everything has some mental or experiential properties, and I remain unconvinced, but I was genuinely interested in hearing how a panpsychist would define consciousness. Those of us who believe that the mind arises from the action of neurons and associated bits of brain tissue have the advantage of a few clues about the kind of thing consciousness is and the kind of thing it certainly isn't; panpsychists, who believe awareness of some kind is a fundamental property of everything, seem to me to have less to go on and consequently a more difficult task in defining exactly what it is. But sometimes it's coming at a problem from a strange new angle that yields useful insights.'

Skrbina very briefly puts a case for panpsychism by noting that even rocks maintain their own existence with a degree of success and respond to the impacts and changes of their environment. This amounts, he suggests, to at least a simple form of experience, and hence of mind. But mind, he says, has two aspects: the inner phenomenal experience and an outward-facing intentional/relational aspect. Both of these are characteristic of the mental life of all things; he acknowledges at least a prima facie difficulty over what counts as a 'thing' here, but it includes such entities as atoms, rocks, tables, chairs, human beings, planets, and stars. In a footnote, Skrbina cites Plato and Aristotle as allies in thinking that stars might have a mental life, together with JBS Haldane's view that the interior of stars might shelter minds superior to our own (perhaps not quite the same view - the existence of minds within stars doesn't imply that the stars themselves have minds any more than the existence of minds in France suggests that France has its own mentality) and Roger Penrose who apparently has speculated that neutron stars may sustain large quantum superpositions and thus conceivably a high intensity of consciousness.

Skrbina does not, of course, believe that rocks have minds exactly like our own, and suggests that material complexity corresponds with mental complexity, so that there is a spectrum of mental life from the feeble, unremembered glimmerings experienced by rocks all the way up to the fantastically elaborate and persistent mental evolutions hosted by human beings. This is convenient, since it allows Skrbina to find a place for subconscious and unconscious mental activity, which can be regarded as merely low-wattage mentality, whereas on the face of it panpsychism seems to make unconsciousness impossible. But, he says, there is a fundamental continuity, and this applies to consciousness as well as general mentality. Consciousness, he suggests, is the border, the interface between the inward and outward aspects of mentality, and since everything possesses both of those everything must have at least a simple analogue of consciousness. It might be better, he suggests, if we could find a new word for this.

Skrbina's exposition is brief, and he only claims to be providing a pointer toward a promising line of investigation. The idea of consciousness as the linkage or interface between inner and outer mentality does have some appeal. Skrbina's distinction between inner and outer corresponds approximately to a view which is widely popular about there being two basic kinds of consciousness: the phenomenal, experiential variety and the rest. Famously this kind of distinction is embodied in David Chalmers' hard/easy problem distinction and Ned Block's a-consciousness and p-consciousness, to name only two examples; the pieces in the JCS provide other variations. Why not regard consciousness as the thing that brings them together, even if you're not attracted by panpsychism?

Well, I don't know. For one thing I think the non-phenomenal half of the mind is usually short-changed. Besides phenomenal awareness, we ought also to distinguish between agency, intentionality, and understanding, all large mysteries which really deserve better than being smooshed together. We could still see consciousness as the thing that brings it all together, perhaps, but that doesn't exactly appeal either: it seems too much like saying that the human body is the thing that holds our bones and muscles together; better to say it's the thing they help to make up.

I must confess - and this perhaps is unfair - to being put off by Skrbina's description of consciousness as the luminous upper layer of the mind. Apart from the slightly confusing geometry (it's the upper layer of the mind, but between the inner and outer parts), I don't see why it's luminous, and that sounds a bit like the resort to poetry sometimes adopted by theologians who have run out of cogent points to make. Still, he deserves at least a couple of cheers for offering a new approach, something he rightly advocates.

How Wrong Can I Be?

14 June 2009

I was pondering the question of my own infallibility recently. Not as the result of a sudden descent into megalomaniacal delusion - I was thinking only of the kinds of infallibility which, if they exist, are shared by all of us conscious beings.

Of course, I am only infallible on certain points: the question is, which? One prime candidate is my own existence. Quite a few people these days contend that the 'self' is an illusion, perhaps a fading shadow of the idea of the soul, or kind of trick with smoke and mirrors. If we are to believe Descartes and his cogito, however, my own existence is the one thing I can't doubt. Non-existent people don't doubt anything, so to doubt my own existence is to prove it - though some would be quick to point out that a momentary doubt doesn't amount to all that much in the way of a self, and that everything else remains to be argued for.

There are, in any case, some other issues on which I may be infallible. I might be infallible about certain aspects of my current experience. I could certainly be mistaken (dreaming, deluded, illuded) about the fact that this is a dagger I see before me, but I couldn't be wrong about the fact that it seems to me there's a dagger before me, could I? The world might be an illusion, but even illusions are things.

Or to make it even less debatable, I could be wrong about having stubbed my toe just now, but I couldn't be wrong about the sensation of pain I'm currently feeling, surely? Perhaps I never stubbed my toe; perhaps I don't have toes and am just a brain in a vat somewhere; perhaps none of the world really exists at all and this current thought, with all its implied memories and feelings, is just a weird metaphysical wrinkle on the surface of universal nothingness; but that experience of pain is finally undeniable. The sheer immediacy of pain seems to mean that there just is no gap between me and it into which any misunderstanding could creep.

And yet… We're familiar these days with the strange deformations of awareness which can result from brain injury; people who no longer recognise their arm as belonging to them; people who feel pain in an arm they haven't got any more; people who are blind but insist, in the teeth of the evidence, that they can see. Isn't it possible we could have people who believe themselves to be in pain when they aren't? A competent hypnotist produces false and even absurd beliefs in subjects all the time and could probably induce such a state without the least difficulty?

Well, a hypnotist could certainly induce someone to say they were in pain, and behave as if they were in pain; but would the subject be in real pain? Unfortunately, the only way we can get at people's real, inner, subjective states is through their reports, so if a hypnotist has interfered with their ability to report, we're a bit stuck. These days, it's true, we could put someone in a scanner and have a look at their brain activity; but that would still beg some philosophical questions.

It's tempting to say, look, I have real pain in my toe right this minute and that - that - can't be a mistake. I grant you could fool some person into declaring themselves in pain falsely, and even

believing it. We could imagine Mary the Pain Scientist, who has lived since birth in a state of analgesia; then we tickle her toes and tell her that that's pain. Of course she believes it. But these cases of error somehow just don't touch the infallibility of the real cases, like mine. Mary, and the other deluded pain-claimants, are simply using the wrong words - they're calling something pain which isn't really pain. But let's put words aside; that thing I'm feeling now - that's what I'm feeling, and I can't be wrong.

If that argument succeeds, it seems to do so only by descending to a level where the concepts of truth and falsity no longer apply: of course there's a sense in which a mere wordless sensation can't be false. It can still be real, but if the reality of my feelings is all we've established, we don't seem to have added anything of substance to the cogito. And indeed if we put ourselves back into Descartes shoes, it seems impossible to deny that some wicked demon could have convinced us that we were in pain when we aren't - that's more or less what happened to Pain Mary.

This is murky territory, but my own guess is that while we can't feel pain without feeling pain, we can believe we're in pain without feeling real pain, and for that matter we can feel pain while holding at some level the erroneous belief that we're not feeling pain. The feeling itself may be veridical in some loose sense, but it can coexist with a higher-order belief about it which happens to be false.

I feel reasonably happy about that because it is clearly the case that we can have false beliefs about our own beliefs; indeed, it's pretty common. I absolutely believe that bungee jumping is safe, until I step onto the platform, when I find that some part of my brain, or perhaps just my legs and stomach, hold other views entirely.

But the mention of believing something with one's legs reveals that beliefs are slippery and polymorphous. To believe something I don't need to do anything; I can have beliefs about things I never thought of (yesterday I believed that Kubla Khan's smallest horse never used the St Malo ferry, and would have said so confidently and without hesitation if asked, though I never thought of the beast until now), and I can go on believing things while unconscious, perhaps indeed when dead (did Galileo stop believing that the Earth goes round the Sun when he died?) .

But what about good old straightforward thoughts? Surely my thought that I am thinking that A, is much less vulnerable than a belief that I believe that A? How could I be thinking about a nice cup of tea while thinking that I'm thinking about the ferry to St Malo? It's certainly possible for the attention to wander from one topic to another by insensible degrees - but could I really be mistaken about what I'm thinking now? There seems a real problem in that to me.

Now of course introspection is systematically unreliable in dealing with questions of this sort. Since introspective thoughts are pretty much by definition second order - ie, they're thoughts about thoughts - introspection only gives me access to half of the comparison. I can only think about thoughts I'm thinking about. If my real thoughts were not what I thought I was thinking, how would I know any different?

It's a fair point, but if my thoughts could be different from what I thought I was thinking, it would surely give rise to some very odd discrepancies between my behaviour and my expectations. In the case of beliefs, we've already noticed that discrepancies of a sort can arise, causing some minor inconsistencies in my behaviour over bungee-jumping, for example, as I stride confidently to the platform and then subside into panicky paralysis - but current thoughts seem a different and more difficult case to me.

How could that be so, though? Thoughts about my thoughts are second-order thoughts, and there is no magic connection between a thought and its target which guarantees accuracy. It follows that there is simply no way of guaranteeing that second-order thoughts are not erroneous, and so there seems to be no way the infallibility I'm attributing to my own thoughts could arise.

The only answer I can see is that we must be wrong to assume that my knowledge of my own thoughts comes from second-order thoughts. The reason I know what I'm thinking is not that I have another true thought about what I'm thinking; instead, my knowledge just comes with the fact that this is what I'm thinking.

That radically contradicts the theory championed by some that conscious thoughts are exactly those for which there is a second order thought about the first thought. The error here, perhaps, has been to construct a model for our knowledge of our own thoughts which resembles our knowledge of the contents of a book. We know what that sentence says because we have a correct thought about it. But books have only derived intentionality (they only mean anything because someone has interpreted them as meaning something) whereas our thoughts have original intentionality (they mean stuff irrespective of what anyone thinks about it). It seems, then, that the distinguishing property of a conscious thought is actually that the having of content and the knowing of content are inseparably the same.

I feel we're frustratingly close to a new insight into the nature of intrinsic intentionality here. But I could be wrong.

Dancing Pixies

1 July 2009

I see that among the first papers published by the recently launched *Journal of Cognitive Computation*, they sportingly included one arguing that we shouldn't be seeing cognition as computational at all. The paper, by John Mark Bishop of Goldsmith's, reviews some of the anti-computational arguments and suggests we should think of cognitive processes in terms of communication and interaction instead.

The first two objections to computation are in essence those of Penrose and Searle, and both have been pretty thoroughly chewed over in previous discussions in many places; the first suggests that human cognition does not suffer the Gödelian limitations under which formal systems must labour, and so the brain cannot be operating under a formal system like a computer program; the second is the famous Chinese Room thought experiment. Neither objection is universally accepted, to put it mildly, and I'm not sure that Bishop is saying he accepts them unreservedly himself - he seems to feel that having these popular counterarguments in play is enough of a hit against computationalism in itself to make us want to look elsewhere.

The third case against computationalism is the pixies: I believe this is an argument of Bishop's own, dating back a few years, though he scrupulously credits some of the essential ideas to Putnam and others. A remarkable feature of the argument is that uses panpsychism in a reductio ad absurdum (reductio ad absurdum is where you assume the truth of the thing you're arguing against, and then show that it leads to an absurd, preferably self-contradictory conclusion).

Very briefly, it goes something like this; if computationalism is true, then anything with the right computational properties has true consciousness (Bishop specifies Ned Block's p-consciousness, phenomenal consciousness, real something-that-it-is-like experience). But a computation is just a given series of states, and those states can be indicated any way we choose. It follows that on some interpretation, the required kind of series of states are all over the place all the time. If that were true, consciousness would be ubiquitous, and panpsychism would be true (a state of affairs which Bishop represents as being akin to a world full of pixies dancing everywhere). But since, says Bishop, we know that panpsychism is just ridiculous, that must be wrong, and it follows that our initial assumption was incorrect; computationalism is false after all.

There are of course plenty of people who would not accept this at all and would instead see the whole argument as just another reason to think that panpsychism might be true after all. Bishop does not spend much time on explaining why he thinks panpsychism is unacceptable, beyond suggesting that it is incompatible with the 'widespread' desire to explain everything in physical terms, but he does take on some other objections more explicitly. These mostly express different kinds of uneasiness about the idea that an arbitrary selection of things could properly constitute a computation with the right properties to generate consciousness.

One of the more difficult is an objection from Hofstadter that the required sequences of states can only be established after the fact: perhaps we could copy down the states of a conscious

experience and then reproduce them, but not determine them in advance. Bishop uses an argument based on running the same consciousness program on a robot twice; the first time we didn't know how it would turn out; the second time we did (because it's an identical robot and identical program) but it's absurd to think that one run could be conscious and the other not.

Perhaps the most tricky objection mentioned is from Chalmers; it points out that cognitive processes are not pre-ordained linear sequences of states, but at every stage have the possibility of branching off and developing differently. We could, of course remove every conditional switch in a given sequence of conscious cognition and replace it by a non-conditional one leading on to the state which was in fact the next one chosen. For that given sequence, the outputs are the same - but we're not entitled to presume that conscious experience would arise in the same way because the functional organisation is clearly different, and that is the thing, on computationalist reasoning, which needs to be the same. Bishop therefore imagines a more refined version: two robots run similar programs; one program has been put through a code optimiser which keeps all the conditional branches but removes bits of code which follow, as it were, the unused branches of the conditionals. Now surely everything relevant is the same: are we going to say that consciousness arises in one robot by virtue of there being bits of extra code there which lie there idle? That seems odd.

That argument might work, but we must remember that Bishop's reductio requires the basics of consciousness to be lying around all over the place, instantiated by chance in all sorts of things. While we were dealing with mere sequences of states, that might look plausible, but if we have to have conditional branches connecting the states (even ones whose unused ends have been pruned) it no longer seems plausible to me. So in patching up his case to respond to the objection, Bishop seems to me to have pulled out some of the foundations it was originally standing on. In fact, I think that consciousness requires the right kind of causal relations between mental states, so that arbitrary sets or lists of states won't do.

The next part of the discussion is interesting. In many ways computationalism looks like a productive strategy, concedes Bishop - but there are reasons to think it has its limitations. One of the arguments he quotes here is the Searlian point that there is a difference between a simulation and reality. If we simulate a rainstorm on a computer, no-one expects to get wet; so if we simulate the brain, why should we expect consciousness? Now the distinction between a simulation and the real thing is a relevant and useful one, but the comparison of rain and consciousness begs the question too much to serve as an argument. By choosing rain as the item to be simulated, we pick something whose physical composition is (in some sense) essential; if it isn't made of water, it isn't rain. To assume that the physical substrate is equally essential for consciousness is just to assume what computationalism explicitly denies. Take a different example, a letter. When I write a letter on my PC, I don't regard it as a mere simulation, even though no paper need be involved until it's printed; in fact, I have more than once written letters which were eventually sent as email attachments and never achieved physical form. This is surely because with a letter, the information is more essential than the physical instantiation. Doesn't it seem highly plausible that the same might be true to an even greater extent of consciousness? If it is true, then the distinction between simulation and reality ceases to apply.

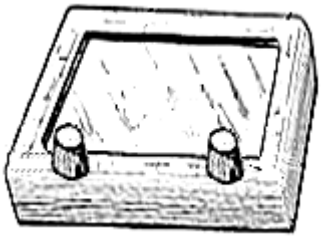
To make sceptical simulation arguments work, we need a separate reason to think that some computation was more like a simulation than the reality - and the strange thing is, I think that's more or less what the objections from Hofstadter and Chalmers were giving us; they both sort of draw on the intuition that a sequence of states could only simulate consciousness in the sort of way a series of film frames simulates motion.

The ultimate point, for Bishop, is to suggest we should move on from the 'metaphor' of computation to another based on communication. It's true that the idea of computation as the basis of consciousness has run into problems over recent years, and the optimism of its adherents has been qualified by experience; on the other hand it still has some remarkable strengths. For one thing, we understand computation pretty clearly and thoroughly; if we could reduce consciousness to computation, the job really would be done; whereas if we reduce consciousness to some notion of communication which still (as Bishop says) requires development and clarification, we may still have most of the explanatory job to do.

The other thing is that computation of some kind, if not the only game in town, still is far closer to offering a complete answer than any other hypothesis. I suspect many people who started out in opposing camps on this issue would agree now that the story of consciousness is more likely to be 'computation plus plus' (whatever that implies) than something quite unrelated.

Chaotic Consciousness

14 July 2009



An interesting New Scientist piece recently reviewed research suggesting that chaos has an important part in the way the brain functions. More specifically, the suggestion is that the brain operates 'on the edge of chaos', in self-organised criticality; sometimes it runs in ways which are predictable at a macro level, more or less like a conventional machine; but at times it also goes into chaotic states. The behaviour of the system in these states is still fully deterministic in a wholly traditional, classical way, but depends so exquisitely on the fine detail of the starting state that the behaviour of the system is in practice unpredictable. The analogy offered here is a growing pile of sand; you can't tell exactly when it will suddenly go through a state shift - collapse - although over a long period the number of large and small collapses is amenable to statistical treatment (actually, I have to say I've never noticed piles of sand behaving in this interesting way, but that just shows what a poor observer I am).

The suggestion is that the occasional 'avalanches' of neuronal firing in the brain are useful, allowing the brain to enter new states more rapidly than it could otherwise do. Being on the edge of chaos allows "maximum transmission with minimum risk of descending into chaos". The arrival of a neuronal avalanche is related to the sudden popping-up of an idea in the mind, or perhaps the unexpected recurrence of a random memory. There is also evidence that the duration of phase-shifts is related to IQ scores – perhaps in this case because the longer shift allows the recruitment of more neurons. The recruitment of additional neurons is presumed in such cases to be a good thing (I feel there must be some caveats about that), but there are also suggestions that excess time spent in phase-shifts could be a cause of schizophrenia (someone should set out a list somewhere of all the things that at one time or another have been put forward as causes of schizophrenia); while not enough phase-shifting in parts of the brain to do with social behaviour might have something to do with autism.

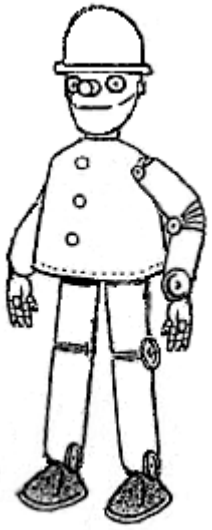
One claim not made in the article, but one which could well be made, is that all this might account for the sensation of free will. If the brain occasionally morphs through chaos into a new state, might it not be that the conclusions which emerge would seem to have come out of nowhere? We might be led to assume that these thoughts were freely generated, distinct from the normal predictable pattern. I think the temptation would be to frame such a theory as an explanation of the illusion of free will: why we feel as if some of our decisions are free even though, in the final analysis, determinism rules. But I can also imagine that a compatibilist might claim that chaotic phase shifts really were freedom. A free act is one which is not predictable, such a person might argue; however, we don't mean unpredictable in practice – none of us is currently able to look at a brain and predict the decisions it will make in any given circumstances. We mean predictable in principle; predictable if we had all the data plus unlimited time and computing power. Now are chaotic changes predictable in principle or not? They occur within normal physical rules, so in the ultimate sense they are clearly deterministic. But the difficulties are so great that to say that they're only unpredictable in practice seems to stretch 'practice' a long way – we might easily need perfection of measurement to a degree which is never going to be obtainable under any imaginable real circumstances. Couldn't we rather say, then, that we're dealing with a third kind of unpredictability, neither quite unpredictability in mere practice nor quite unpredictability in principle, and take the view that

decisions subject to this level of unpredictability deserve to be called free? I think we could, but ultimately I'm disinclined to do so because in the final analysis that feels more like inventing a new concept of freedom than justifying the existing one.

There's another issue here that affects a number of the speculations in the article. We must beware of assuming too easily that features of the underlying process necessarily correspond directly with phenomenal features of experience. So, for example, it's assumed that when the brain goes quickly into a new state in terms of its neuronal firing, that would be like a new thought popping up suddenly in our conscious minds, an idea which seemed to have come out of nowhere. It ain't necessarily so (though it would be an interesting question to test). The fact that the brain uses chaos to achieve its results does not mean that the same chaos is directly experienced in our thoughts, any more than I experience say, that old 40Hz buzz starting up in my right parietal, or whatever. At the moment (not having read the actual research, of course) it seems equally likely that phase shifts are wholly outside conscious experience, perhaps, for example, being required in order to allow subordinate systems to catch up rapidly with a separate conscious process which they don't directly influence. Or perhaps they're just the vigorous shaking which clears our mental etch-a-sketch, correlated with but not constitutive of, the sophisticated complication of our conscious doodlings.

The Three Laws Revisited

2 August 2009



Ages ago (gosh, it was nearly five years ago) I had a piece where Blandula remarked that any robot clever enough to understand Isaac Asimov's Three Laws of Robotics would surely be clever enough to circumvent them. At the time I think all I had in mind was the ease with which a clever robot would be able to devise some rationalisation of the harm or disobedience it was contemplating. Asimov himself was of course well aware of the possibility of this kind of thing in a general way. Somewhere (working from memory) I think he explains that it was necessary to specify that robots may not, through inaction, allow a human to come to harm, or they would be able to work round the ban on outright harming by, for example, dropping a heavy weight on a human's head. Dropping the weight would not amount to harming the human because the robot was more than capable of catching it again before the moment of contact. But once the weight was falling, a robot without the additional specification would be under no obligation to do the actual catching.

That does not actually wrap up the problem altogether. Even in the case of robots with the additional specification, we can imagine that ways to drop the fatal weight might be found. Suppose, for example, that three robots, who in this case are incapable of catching the weight once dropped, all hold on to it and agree to let go at the same moment. Each individual can feel guiltless because if the other two had held on, the weight would not have dropped. Reasoning of this kind is not at all alien to the human mind; compare the planned dispersal of responsibility embodied in a firing squad.

Anyway, that's all very well, but I think there may well be a deeper argument here: perhaps the cognitive capacity required to understand and apply the Three Laws is actually incompatible with a cognitive set-up that guarantees obedience.

There are two problems for our Asimovian robot: first it has to understand the Laws; second, it has to be able to work out what actions will deliver results compatible with them. Understanding, to begin with, is an intractable problem. We know from Quine that every sentence has an endless number of possible interpretations; humans effortlessly pick out the one that makes sense, or at least a small set of alternatives; but there doesn't seem to be any viable algorithm for picking through the list of interpretations. We can build in hard-wired input-output responses, but when we're talking about concepts as general and debatable as 'harm', that's really not enough. If we have a robot in a factory, we can ensure that if it detects an unexpected source of heat and movement of the kind a human would generate, it should stop thrashing its painting arm around - but that's nothing like intelligent obedience of a general law against harm.

But even if we can get the robot to understand the Laws, there's an equally grave problem involved in making it choose what to do. We run into the frame problem (in its wider, Dennettian form). This is, very briefly, the problem that arises from tracking changes in the real world. For a robot to keep track of everything that changes (and everything which stays the same, which is also necessary) involves an unmanageable explosion of data. Humans somehow pick out just

relevant changes; but again a robot can only pick out what's relevant by sorting through everything that might be relevant, which leads straight back to the same kind of problem with indefinitely large amounts of data.

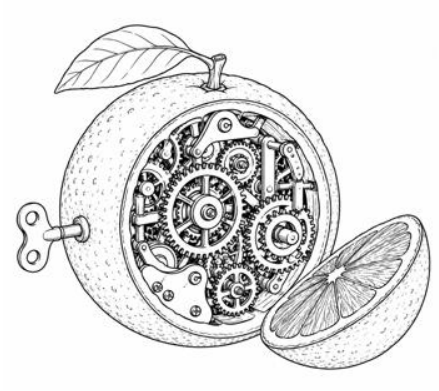
I don't think it's a huge leap to see something in common between the two problems; I think we could say that they both arise from an underlying difficulty in dealing with relevance in the face of the buzzing complexity of reality. My own view is that humans get round this problem through recognition; roughly speaking, instead of looking at every object individually to determine whether it's square, we throw everything into a sort of sieve with holes that only let square things drop through. But whether or not that's right and putting aside the question of how you would go about building such a faculty into a robot, I suggest that both understanding and obedience involve the ability to pick out a cogent, non-random option from an infinite range of possibilities. We could call this free will if we were so inclined, but let's just call it a faculty of choice.

Now I think that faculty, which the robot is going to have to exercise in order to obey the Laws, would also unavoidably give it the ability to choose whether to obey them or not. To have the faculty of choice, it has to be able to range over an unlimited set of options, whereas constraining it to any given set of outcomes involves setting limits. I suppose we could put this in a more old-fashioned mentalistic kind of way by observing that obedience, properly understood, does not eliminate the individual will but on the contrary requires it to be exercised in the right way.

If that's true (and I do realise that the above is hardly a tight knock-down argument) it would give Christians a neat explanation of why God could not have made us all good in the first place - though it would not help with the related problem of why we are exposed to widely varying levels of temptation and opportunity. To the rest of us it offers, if we want it, another possible compatibilist formulation of the nature of free will.

Clockwork Orange Syndrome

29 August 2009



Over at the Institute for Ethics and Emerging Technologies, Martine Rothblatt offers to solve the Hard Problem for us in her piece "Can Consciousness be created in Software?". The Hard Problem, as regulars here will know, is how to explain subjective experience, the redness of red, the ineffable inner sense of what experience is like. Rothblatt's account is quite short, and on my first reading I had slipped past the crucial sentences and discovered she was claiming victory before I quite realised what her answer was, so that I had to go back and read it again more carefully.

She says: "Redness is part of the gestalt impression obtained in a second or less from the immense pattern of neural connections we have built up about red things... With each neuron able to make as many as 10,000 connections, and with 100 billion neurons, there is ample possibility for each person to have subjective experiences through idiosyncratic patterns of connectivity. The point seems to be that the huge connectivity of the human brain makes it easy for everyone's experience of redness to be unique.

This is an interesting and rather novel point of view. One traditional way of framing the Hard Problem is to ask whether the blue that I see is really the same as the blue that you see - or could it be equivalent to my orange? How would we know? On Rothblatt's view it would seem the answer must be that my blue is definitely not the same as yours, nor the same as anyone else's; and nor is it the same as your orange, or Fred's green for that matter, everyone having an experience of blue which is unique to them and their particular pattern of neuronal connection. I find this a slightly scary perspective, but not illogical. Presumably from Rothblatt's point of view the only thing the different experiences have in common is that they all refer to the same blue things in the real world (or something like that).

Is it necessarily so, though? Does the fact that our neuronal connections are different mean our experiences are different? I don't think so. After all, I can write a particular sentence in different fonts and different sizes and colours of text; I can model it in clay or project it on a screen, yet it remains exactly the same sentence. Differences in the substrate don't matter. We can go a step further: when people think of that same sentence, we must presume that the neuronal connections supporting the thought are different in each individual - yet it's the same sentence they're thinking. So why can't different sets of neural connections support the same experience of blue in different heads?

Rothblatt's account is a brief one, and perhaps she has some further theoretical points which explain the relationship between neuronal structures and subjective experiences in a way which would illuminate this. But hang on - wasn't the relationship between neuronal structures and subjective experience the original problem? Slapping our foreheads, we realise that Rothblatt has palmed off on us an idea about uniqueness which actually doesn't address the Hard Problem at all. The Hard Problem is not about how brains can accommodate a million unique experiences: it's about how brains can accommodate subjectivity, how they make it true that there is something it is like to see red.

I fear this may be another example of the noticeable tendency of some theorists, especially those with a strong scientific background, to give us clockwork oranges, to borrow the metaphor Anthony Burgess used in a slightly different context: accounts that model some of the physical features quite nicely (see, it's round, it hangs from trees effectively) while missing the essential juice which is really the whole point. According to Burgess, or at least to F. Alexander, the ill-fated character who expresses this view in the novel, the point of a man is not orderly compliance with social requirements but, as it were, his juice, his inner reality. I think the answer to the Hard Problem is indeed something to do with our reality (I should say that I mean 'reality' in a perfectly everyday sense: Rothblatt, like some purveyors of mechanical fruit, is a little too quick to dismiss anything not strictly reducible to physics as mysticism) : we're fine when we're dealing with abstractions, it's the concrete and particular which remains stubbornly inexplicable in any but superficial terms. It's my real experience of redness that causes the difficulty. Perhaps Rothblatt's ideas about uniqueness are an attempt to capture the one-off quality of that experience; but uniqueness isn't quite the point, and I think we need something deeper.

Bafflement

19 September 2009

Over at Robots.net, they've noticed a bit of a resurgence of dualism recently, and it seems that Avshalom C. Elitzur is in the vanguard, with this paper presenting an argument from bafflement.

The first part of the paper provides a nice, gentle introduction to the issue of qualia in dialogue form. Elitzur explains the bind that we're in in this respect: we seem to have an undeniable first-hand experience of qualia, yet they don't fit into the normal physical account of the world. We seem to be faced with a dilemma: either reject qualia - perhaps we just misperceive our percepts as qualia - or accept some violation of normal physics. The position is baffling: but Elitzur wants to suggest that that very bafflement provides a clue. His strategy is to try to drag the issue into the realm of science, and the argument goes like this:

1. By physicalism, consciousness and brain processes are identical.
2. Whence, then, the dualistic bafflement about their apparent nonidentity?
3. By physicalism, this nonidentity, and hence the resultant bafflement, must be due to error.
4. But then, again by physicalism, an error must have a causal explanation.
5. Logic, cognitive science and AI are advanced enough nowadays to provide such an explanation for the alleged error underlying dualism, and future neurophysiology must be able to point out its neural correlate.

That last point seems optimistic. Cognitive science may be advanced enough to provide explanations for a number of cognitive deficits and illusions, but sometimes only partial ones; and not all errors are the result of a structural problem. It's particularly optimistic to think that all errors must have an identifiable neural correlate. But this seems to be what Elitzur believes. He actually says

"When future neurophysiology becomes advanced enough to point out the neural correlates of false beliefs, a specific correlate of this kind would be found to underlie the bafflement about qualia."

The neural correlates of false beliefs? Crikey! It's perfectly reasonable to assume that all false beliefs have neural correlates - because one assumes that all beliefs do - but the idea that false ones can be distinguished by their neural properties is surely evidently wrong. An argument hardly seems required, but it's easy, for example, to picture a man who believes a coin has come down heads. If it has, his belief is true, but if it's actually tails, exactly the same belief, with identical neural patterns would be false. I think Elitzur must mean something less startling than what he seems to be saying; he must, I think, take it as read that if qualia are a delusion, they would be a product of some twist or quirk in our mental set-up. That's not an unreasonable position, one that would be shared by Metzinger, for example (discussion coming soon).

As it happens, Elitzur doesn't think qualia are delusions; instead he has an argument which he thinks shows that interactionist dualism - a position he doesn't otherwise find very attractive - must be true. The argument is to do with zombies. Zombies in this context, as regular readers

will know, are people who have all the qualities normal people posess, except qualia. Because qualia have no physical causal effects, the behaviour of zombies, caused by normal physical factors, is exactly like that of normal people. Elitzur quotes Chalmers explaining that zombie-Chalmers even talks about qualia and writes philosophical papers about them, though in fact he has none. The core of Elitzur's position is his incredulity over this conclusion. How could zombies who don't have qualia come to be worried about them?

It is an uncomfortable position, but if we accept that zombies are possible and qualia exist, Chalmers' logic seems irrefutable. Ex hypothesi, zombies follow the same physical laws as us: it's ultimately physics that causes the movements of our hands and mouths involved in writing or speaking about qualia: so our zombie counterparts must go through the same motions, writing the same books and emitting the same sounds. Since this seems totally illogical to Elitzur, he offers the rationalisation that when zombies talk about qualia, they must in fact merely be talking about their percepts. But this asymmetry provides a chink which can be used to prose zombies and qualiate people apart. If we ask Chalmers whether his zombie equivalent is possible, he replies that it is; but, suggests Elitzur, if we ask zombie Chalmers (whom he call 'Charmless') the same question, he replies in the negative. Chalmers can imagine himself functioning without qualia, because qualia have no functional role: but Charmless cannot imagine himself functioning without percepts, because percepts are part of the essence of his sensory system. (It is possible to take the analogous view about qualia of course - namely that zombies are impossible, because a physically identical person just would necessarily have the same qualia). So zombies differ from us, oddly enough, in not being able to conceive of their own zombies.

For Elitzur, the conclusion is inescapable; qualia do have an effect on our brains. He chooses therefore to bite the bullet of accepting that the laws of physics must be messed up in some way - that where qualia intervene, conservation laws are breached, unpalatable as this conclusion is. One consoling feature is that if qualia do have physical effects, they can be included in the evolutionary story; perhaps they serve to hasten or intensify our responses: but overall it's regrettable that dualism turns out to be the answer.

I don't think this is a convincing conclusion; it seems as if Elitzur's incredulity has led him into not taking the premises of the zombie question seriously enough. It just is the case ex hypothesi that all of our zombies' behaviour is caused by the same physical factors as our own behaviour; it follows that if their talk about qualia is not caused by qualia, neither is ours (note that this doesn't have to mean that either we or the zombies fail to talk about qualia). There are other ways out of this uncomfortable position, discussed by Chalmers (perhaps, for example, our words about qualia are over-determined, caused both by physical factors and by our actual experiences). My own preferred view is that whatever qualia might be, they certainly go along with certain physical brain functions, and that therefore any physical duplicate of ourselves would have the same qualia; that zombies, in other words, are not possible. It's just a coincidence, I'm sure, that in Elitzur's theory this is the kind of thing a zombie would say.

Libet was wrong..?

26 September 2009

One of the most frequently visited pages on Conscious Entities is an account of Benjamin Libet's remarkable experiments, which seemed to show that decisions to move were really made half a second before we were aware of having decided. To some this seemed like a practical disproof of the freedom of the will - if the decision was already made before we were consciously aware of it, how could our conscious thoughts have determined what the decision was? Libet's findings have remained controversial ever since they were published; they have been attacked from several different angles, but his results were confirmed and repeated by other researchers and seemed solid.

However, Libet's conclusions rested on the use of Readiness Potentials (RPs). Earlier research had shown that the occurrence of an RP in the brain reliably indicated that a movement was coming along just afterwards, and they were therefore seen as a neurological sign that the decision to move had been taken (Libet himself found that the movement could sometimes be suppressed after the RP had appeared, but this possibility, which he referred to as 'free won't ', seemed only to provide an interesting footnote). The new research, by Trevena and Miller at Otago, undermines the idea that RPs indicate a decision.

Two separate sets of similar experiments were carried out. They resembled Libet's original ones in most respects, although computer screens and keyboards replaced Libet's more primitive equipment, and the hand movement took the form of a keypress. A clock face similar to that in Libet's experiments was shown, and they even provided a circling dot. In the earlier experiments this had provided an ingenious way of timing the subject's awareness that a decision had been made - the subject would report the position of the dot at the moment of decision - but in Trevena and Miller's research the clock and dot were provided only to make conditions resemble Libet's as much as possible. Subjects were told to ignore them (which you might think rendered their inclusion pointless). This was because instead of allowing the subject to choose their own time for action, as in Libet's original experiments, the subjects in the new research were prompted by a randomly timed tone. This is obviously a significant change from the original experiment; the reason for doing it this way was that Trevena and Miller wanted to be able to measure occasions when the subject decided not to move as well as those when there was movement. Some of the subjects were told to strike a key whenever the tone sounded, while the rest were asked to do so only about half the time (it was left up to them to select which tones to respond to, though if they seemed to be falling well below a 50-50 split, they got a reminder in the latter part of the experiment). Another significant difference from Libet's tests is that left and right hands were used: in one set of experiments the subjects were told by a letter in the centre of the screen whether they should use the right or left hand on each occasion, in the other it was left up to them.

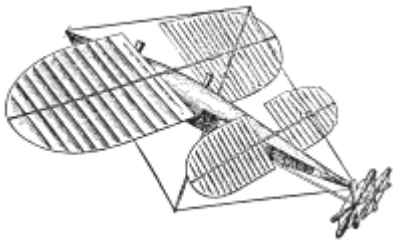
There were two interesting results. One was that the same kind of RP appeared whether the subject pressed a key or not. Trevena and Miller say this shows that the RP was not, after all, an indication of a decision to move, and was presumably instead associated with some more general kind of sustained attention or preparing for a decision. Second, they found that a different kind of RP, the Lateralised Readiness Potential or LRP, which provides an indication of

readiness to move a particular hand, did provide an indication of a decision, appearing only where a movement followed; but the LRP did not appear until just after the tone. This suggests, in contradiction to Libet, that the early stages of action followed the conscious experience of deciding, rather than preceding it.

The differences between these new experiments and Libet's originals provide a weak spot which Libetians will certainly attack. Marcel Brass, whose own work with fMRI scanning confirmed and even extended Libet's delay, seeming to show that decisions could be predicted anything up to ten seconds before conscious awareness, has apparently already said that in his view the changes undermine the conclusions Trevena and Miller would like to draw. Given the complex arguments over the exact significance of timings in Libet's results, I'm sure the new results will prove contentious. However, it does seem as if a significant blow has been struck for the first time against the foundations of Libet's remarkable results.

Cognitive Planes

18 October 2009



I see via MLU that Robert Sloan at the University of Illinois at Chicago has been given half a million dollars for a three-year project on common sense. Alas, the press release gives few details, but Sloan describes the goal of common sense as "the Holy Grail of artificial intelligence research".

I think he's right. There is a fundamental problem here that shows itself in several different forms. One is understanding: computers don't really understand anything, and since translation, for example, requires understanding, they've never been very good at it. They can swap a French word for an English word, but without some understanding of what the original sentence was conveying, this mechanical substitution doesn't work very well. Another outcrop of the same issue is the frame problem: computer programs need explicit data about their surroundings, but updating this data proves to be an unmanageably large problem, because the implications of every new piece of data are potentially infinite. Every time something changes, the program has to check the implications for every other piece of data it is holding; it needs to check the ones that are the same just as much as those that have changed, and the task rapidly mushrooms out of control. Somehow, humans get round this: they seem to be able to pick out the relevant items from a huge background of facts immediately, without having to run through everything.

In formulating the frame problem originally back in the 1960s, John McCarthy speculated that the solution might lie in non-monotonic logics; that is, systems that don't require everything to be simply true or false, as old-fashioned logical calculus does. Systems based on rigid propositional/predicate calculus needed to check everything in their database every time something changed in order to ensure there were no contradictions, since a contradiction is fatal in these formalisations. On the whole, McCarthy's prediction has been borne out in that research since then has tended towards the use of Bayesian methods, which can tolerate contradictions and which can give propositions degrees of belief rather than simply holding them true or false. As well as providing practical solutions to frame problem issues, this seems intuitively much more like the way a human mind works.

Sloan, as I understand it, is very much in this tradition; his earlier published work deals with sophisticated techniques for the manipulation of Horn knowledge bases. I'm afraid I frankly have only a vague idea of what that means, but I imagine it is a pretty good clue to the direction of the new project. Interestingly, the press release suggests the team will be looking at CYC and other long-established projects. These older projects tended to focus on the accumulation of a gigantic database of background knowledge about the world, in the possibly naive belief that once you had enough background information, the thing would start to work. I suppose the combination of unbelievably large databases of common sense knowledge with sophisticated techniques for manipulating and updating knowledge might just be exciting. If you were a cyberpunk fan and unreasonably optimistic, you might think that something like the meeting of Neuromancer and Wintermute was quietly happening.

Let's not get over-excited, though, because of course the whole thing is completely wrong. We may be getting really good at manipulating knowledge bases, but that isn't what the human brain does at all. Or does it? Well, on the one hand, manipulating knowledge bases is all we've

got: it may not work all that well, but for the time being it's pretty much the only game in town - and it's getting better. On the other hand, intuitively it just doesn't seem likely that that's what brains do: it's more as if they used some entirely unknown technique of inference which we just haven't grasped yet. Horn knowledge bases may be good, but really are they any more like natural brain functions than Aristotelian syllogisms?

Maybe, maybe not: perhaps it doesn't matter. I mentioned the comparable issue of translation. Nobody supposes we are anywhere near doing translation by computation in the way the human brain does it, yet the available programs are getting noticeably better. There will always be some level of error in computer translation, but there is no theoretical limit to how far it can be reduced, and at some point it ceases to matter: after all, even human translators get things wrong.

What if the same were true for knowledge management? We could have AI that worked to all intents and purposes as well as the human brain yet worked in a completely different way. There has long been a school of thought that says this doesn't matter: we never learnt to fly the way birds do, but we learnt how to fly. Maybe the only way to artificial consciousness in the end will be the cognitive equivalent of a plane. Is that so bad?

If the half-million dollars is well spent, we could be a little closer to finding out...

The Ego Tunnel

8 November 2009

The denial of one's own existence might seem a desperate philosophical strategy, but denying the reality of the self is a line which a number of people have taken, and Thomas Metzinger is prominent among them. The thesis of his massive 2003 work is summed up in the title: Being No One: The Self-Model Theory of Subjectivity. In that book, Metzinger made a commendable effort to balance philosophy and science; but the sheer size of the resulting text may have deterred some readers – I confess to being somewhat daunted myself. Now he has come back with a slimmer volume The Ego Tunnel which is aimed at a wider public and raises wider issues which Metzinger suggests need public attention.

Metzinger's theory – the Self-model Theory of Subjectivity or SMT - suggests that subjective experience is really a kind of trick the brain plays on itself. Our brain sets up a model of the world (actually based on fairly limited data) to which it then adds a model of us, ourselves. The coherence of the model and the fact that the processes supporting it are transparent – ie invisible to us – yield the vivid impression of a self in direct contact with reality, and that's where subjectivity arises; although in fact the whole thing is simply an illusion.

Metzinger's view of qualia is characteristically complex. He has a good argument against the existence of what he calls canonical qualia, qualia conceived as subjective universals. He points out that our ability to discriminate is far greater than our ability to recognise. So, if we are presented with examples of green 64 and green 66, we can readily tell the difference: but if at a later stage we are presented with one of the examples, we have no hope of telling which it is. So there is no single thing that consistently goes along with the experience of green 64.¶Concluding that at any rate we need to distinguish between 'qualia' available to memory and qualia available to the faculty of recognition, Metzinger goes on to distinguish a series of possible conceptions of qualia, ending with 'Metzinger qualia' which are available attentionally but not cognitively. These are slippery customers for obvious reasons, impossible to report and broadly ineffable – but then that's how qualia are generally assumed to be.

Even as a summary, the foregoing is a bare and radically, probably over- simplified view of the theory, however. Metzinger actually presents ten constraints which need to be satisfied for the occurrence of subjective experience: they are:¶

- Globality; as in 'global workspace'; conscious items are always integrated into an overall world-model,
- Presentationality; present in the now , temporal immediacy,
- Convolved holism; objects of experience are made up of collections of other objects in a nested hierarchy,
- Dynamicity; the perceived world flows through constant changes,
- Perspectivalness; we experience the world from a point of view,

- Transparency; we cannot see the works – the neural processing which gives rise to our experience is excluded from conscious experience,
- Offline activation; subjective experience is not confined to the live inputs from our senses, a notable exception being dreams, where a whole non-existent world appears.
- Representation of intensities; besides distinguishing between qualia (whichever version we're dealing with) we can distinguish their levels of intensity,
- Homogeneity; areas of pink, for example, are made up of smaller areas of pink, not red and white at once, and
- Adaptivity; the features of subjective experience have to be things that could reasonably have appeared in the course of evolution.

You may feel that there's something a bit odd about this list, especially the appearance of 'adaptivity'. One does not have to be a creationist to feel uneasy about the idea that that is really an essential feature of conscious experience in any deep sense. In a précis which he prepared for a discussion on the Psyche site – sadly this no longer appears to be available – Metzinger discussed only the first six constraints, which suggests that the last four are at least somewhat dispensable. This is a bit confusing – is homogeneity essential to conscious experience, or was that just a kind of bonus, a description of a property of phenomenal experience which is important but not a defining requirement? I think one issue here may be that Metzinger seems to want to do two jobs at once; he wants to explain subjectivity in a philosophical sense, but he also wants to describe and categorise phenomenal experience in a way which can be related to scientific observation, clarifying the puzzling phenomenology of blindsight and other unusual conditions. These are not incompatible aims, exactly, and there's no absolute reason why one account couldn't do both, but there is a tension. A robust philosophical explanation would pare down the account to its essentials, while the description of subjectivity seems to be something we would want to do as fully as possible.

So Metzinger, for example, notes that convolution is one property of phenomenal experience and homogeneity is another. If we're looking for one kind of explanation, to be told that subjective experience breaks down into things which are different, and also that subjective experience sometimes breaks down into things that are the same, is not very helpful; but as a description of different ways phenomenology can work it's a legitimate clarification.

In the new book, Metzinger boils the requirement down in terms which make it easier for us to get to grips with the theory in ways that relate to our usual concerns here, by asking what it would take to build an Ego Machine, an artificial conscious subject.

The recipe requires us to begin with an integrated and dynamical world-model . We must then ensure that the information flow is organised so as to assign events to a present moment, a Now; and that the images presented by the model are transparent: again, we must not be able to see the works. If these constraints are satisfied, a world appears. The last step is to add to the model an equally transparent representation of an experiencing self; then the artificial entity will finally appear to itself to be someone, and to be there. There will seem to be someone in the Ego Tunnel, Metzinger's equivalent of Plato's cave.

This seems to me to make quite a lot of sense, but the reader's verdict is likely to be determined by whether it 'feels right' – does it finally dispel that sense of mystery? It must be taken into account that Metzinger's theory itself provides strong reasons why his conclusions are going to seem intuitively unacceptable, so a kind of mental 'aiming off' may be required. I think we have to ask ourselves two questions. The first is: can we imagine that all of Metzinger's requirements

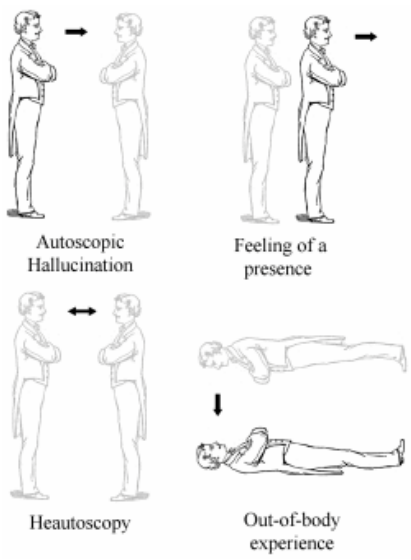
could be met in the absence of 'inner experience': does it seem that we could have his kind of world-model running with a representation of a self in it and yet still be, as it were, only a Metzinger zombie?

If we're happy that this is not a possibility, there still remains the question of whether Metzinger has smuggled his consciousness into the tunnel. Do any of his constraints covertly imply that consciousness is already present?

I think some of them could – presentationality? Perspectivalness? - but carefully read, I don't think they need do. Whether some of their intuitive appeal derives from the reader having unconsciously done some illicit importation is, I suppose, a matter for the reader. Some of the constraints do look a bit strange in other ways. Let's consider that requirement of transparency, for example. In the ego tunnel, data about the outside world must be globally available, but there must be no information about how this data got to the tunnel, since that would dispel the illusion that what we are dealing with is the real world itself. But why would there be? A TV screen does not automatically contain an account of how the TV itself or the local transmitter works. Metzinger has an odd perspective here which seems to start with the idea that the mind naturally knows everything and has to be carefully shielded. In The Ego Tunnel he talks about our sensory apparatus filtering out most of the potential information about the world, limiting us to a meagre trickle, as though our natural state was a kind of omniscience which is only reined in by the limitations of our eyes. Isn't it the other way round?

Look at it the other way: suppose we were consciously aware of the fact that light rays were being processed by our retinas, would that destroy our sense of ourselves as existing in the world? Actually, I doubt it. The fact that we can see the edges of the screen does not inhibit us much from getting immersed in a film: for that matter the very visible appearance of the book does not prevent our getting immersed in a novel. Similarly, the fact that I can see a 'floater' in my right eye at the moment doesn't cause reality to recede from me although it is a tell-tale sign of the visual processes that intervene between my brain and the world.

Among a number of interesting features, The Ego Tunnel includes a substantial account of out-of-body experiences (OBEs) and similar phenomena. Experiments where the subjects are tricked into mistaking a plastic dummy for their real hand (all done with mirrors), or into feeling themselves to be situated somewhere behind their own head (you need a camera for this) show that our perception of our own body and our own location are generated within our brain and are susceptible to error and distortion; and according to Metzinger this shows that they are really no more than illusions (Is that right, by the way - or are they only illusions when they're wrong or misleading? The fact that a camera can be made to generate false or misleading pictures doesn't mean that all photographs are delusions, does it?).



There are many interesting details in this account, quite apart from its value as part of the overall argument. Metzinger briefly touches on four varieties of autoscopic (self-seeing) phenomena, all of which can be related to distinct areas of the brain: autoscopic hallucination,

where the subject sees an image of themselves; the feeling of a presence, where the subject has the strong sense of someone there without seeing anyone; the particularly disturbing heautoscopy, where the subject sees another self and switches back and forth into and out of it, unsure which is 'the real me'; and the better-known OBE. OBEs arise in various ways: often detachment from the body is sudden, but in other cases the second self may lift out gradually from the feet, or may exit the corporeal body via the top of the head. Metzinger tells us that he himself has experienced OBEs and made many efforts to have more (going so far as to persuade his anaesthetist to use ketamine on him in advance of an operation, with no result - I wonder whether the anaesthetist actually kept his word) ; speaking of lucid dreams, another personal interest, he tells the story of having one in which he dreamed an OBE. That seems an interesting bit of evidence: if you can dream a credible OBE, mightn't they all be dreams? This seems to undercut the apparently strong sense of reality which typically accompanies them.

Interestingly, Metzinger reports that a conversation with Susan Blackmore helped him understand his own experiences. Blackmore is of course another emphatic denier of the reality of the self. I don't in any way mean to offer an ad hominem argument here, but it is striking that these two people both seem to have had a particular interest in 'spooky' dualistic phenomena which their rational scientific minds ultimately rejected, leading on to an especially robust rejection of the self. Perhaps people who lean towards dualism in their early years develop a particularly strong conception of the self, so that when they adopt monist materialism they reject the self altogether instead of seeking to redefine and accommodate it, as many of us would be inclined to do?

On that basis, you would expect Metzinger to be the hardest of hard determinists; his ideas seem to lean in that direction, but not decisively. He suggests that certain brain processes involved in preparing actions are brought up into the Ego Tunnel and hence seem to belong to us. They seem to be our own thoughts, our own goals and because the earlier stages remain outside the Tunnel, they seem to have come from nowhere, to be our own spontaneous creations. There are really no such things as goals in the world, any more than colours, but the delusion that they do exist is useful to us; the idea of being responsible for our own actions enables a kind of moral competition which is ultimately to our advantage (I'm not quite sure exactly how this works). But in this case Metzinger pulls his punch: perhaps this is not the full story, he says, and describes compatibilism as the most beautiful position.

Metzinger pours scorn on the idea that we must have freedom of the will because we feel our actions to be free, yet he does give an important place to the phenomenology of the issue, pointing out that it is more complex than might appear. The more you look at them, he suggests, the more evasive conscious intentions become. How curious it is then, that Metzinger, whose attention to phenomenology is outstandingly meticulous, should seem so sure that we have at all times a robust (albeit delusional) sense of our selves. I don't find it so at all, and of course on this no less a person than David Hume is with me; with characteristically gentle but devastating scepticism, he famously remarked "For my part, when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and never can observe any thing but the perception."

Metzinger concludes by considering a range of moral and social issues which he thinks we need to address as our understanding of the mind improves. In his view, for example, we ought not to try to generate artificial consciousness. As a conscious entity, the AI would be capable of suffering, and in Metzinger's view the chances are its existence would be more painful than

pleasant. One reason for thinking so is the constrained and curtailed existence it could expect; another is that we only have our own minds to go on and would be likely to produce inferior, messed-up versions of it. But more alarming, Metzinger argues that human life itself involves an overall preponderance of pain over pleasure; he invokes Schopenhauer and Buddha. With characteristic thoroughness, he concedes that pleasure and pain may not be all that life is about; other achievements can justify a life of discomfort. But even so, the chances for an artificial consciousness, he feels are poor.

This is surely too bleak. I see no convincing reason to think that pain outweighs pleasure in general (certainly the Buddhist case, based on the perverse assumption that change is always painful, seems a weak point in that otherwise logical religion), and I see some reasons to think that a conscious robot would be less vulnerable to bad experiences than we are. It's millions of years of evolution which have ingrained in us a fear of death and the motivating experience of pain: the artificial consciousness need have none of that, but would surely be most likely to face its experiences with superhuman equanimity.

Of course caution is justified, but Metzinger in effect wants us to wait until we've sorted out the meaning of life before we get on with living it.

His attempt to raise this and other issues is commendable though; he's right that the implications of recent progress have not received enough intelligent attention. Unfortunately I think the chances of some of these issues being addressed with philosophic rationality are slim. Another topic Metzinger raises, for example, is the question of what kinds of altered or enhanced mental states, from among the greatly expanded repertoire we are likely to have available in the near future, we ought to allow or facilitate; not much chance that his mild suggestions on that will have much impact.

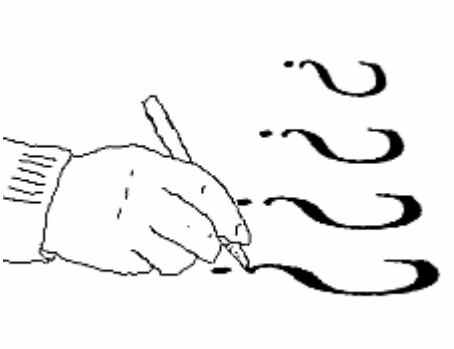
There's a vein of pessimism in his views on another topic. Metzinger fears that the progress of science, before the deeper issues have been sorted out, could inspire an unduly cynical, stripped-down view of human nature; a 'vulgar materialism', he calls it. Uninformed members of the public falling prey to this crude point of view might be tempted to think:

"The cat is out of the bag. We are gene-copying bio-robots, living out here on a lonely planet in a cold and empty physical universe. We have brains but no immortal souls and after seventy years or so the curtain drops. There will never be an afterlife, or any kind of reward or punishment for anyone… I get the message."

Gosh: do we know anyone vulgar and unsophisticated enough to think like that?

Signatures of Consciousness

14 December 2009



Edge has an interesting talk by Stanislas Dehaene. He and his team, using a range of tools, have identified 'signatures' of awareness; marked changes in activity in certain brain regions which accompany awareness of a particular stimulus. They made a clever use of the phenomenon of 'masking', in which the perception of a word can be obliterated if it follows too rapidly after the presentation of an earlier one. By adjusting the relevant delay, the team could compare the effect of a stimulus which never reached consciousness with one that did. Using this and similar techniques they identified a number of clear indications of conscious awareness: increased activity in the early stages of processing and activity in new regions, including the prefrontal cortex and inferior parietal. It appears that this is accompanied by a 'P3 wave' which is quite easy to detect even with nothing more sophisticated than electrodes on the scalp. Interestingly it seems that the difference between a stimulus which does not make it into consciousness and one which does emerges quite late, after as much as a quarter of a second of processing.

No-one, I suspect, is going to be amazed by the news that conscious awareness is accompanied by distinctive patterns of brain activity; but identifying the actual 'signatures' has direct clinical relevance in cases of coma and apparent persistent vegetative state. In principle Dehaene's research should allow conscious reactions continuing in a paralysed patient to be identified; this possibility is being actively pursued.

More speculative and perhaps of deeper theoretical interest, Dehaene puts forward a theory of consciousness as a global neuronal workspace, another variation on the global workspace theory of Bernard Baars (an idea which keeps being picked up by others, which must suggest that it has something going for it). Dehaene offers the view that a particular function of the workspace is to allow inputs to hang around for an extended period instead of dissipating. Among other benefits, this allows the construction of chains of processing operations, something Dehaene likens to a Turing machine, though it sounds a little messier than that to me. Further ingenious experiments have lent support to this idea; the researchers were able to contrast subjects' chaining ability when information was supplied subliminally or consciously (this may sound odd, but subjects can perform at better-than-chance levels even with subliminal stimuli).

Dehaene says that he is dealing only with one variety of consciousness - in the main it's awareness, which in some respects is the basement level compared to the more high-flown self-reflective versions. But in passing the talk does clarify a question which has sometimes troubled me in the past about global workspace theories - why should they involve consciousness at all? It seems easy to understand that the brain might benefit from a kind of clearing house where information from different sources is shared - but couldn't that happen, as it were, in the dark? What does the magic ingredient of consciousness add to the process?

Well, being in the global workspace means being accessible to several different systems (no intention here to commit to any particular view about modularity); and one of those systems is

the vocal reporting system. So as a natural consequence of being in the workspace, inputs become things we can vocally report, things we can talk about. Things we can talk about are surely objects of consciousness in some quite high-level sense.

Dehaene does not go down this path, but I wondered how far we could take it; is there a plausible explanation of phenomenal consciousness in terms of a global workspace? If we followed the same pattern of argument we used above, we would be looking to say that conscious experiences acquired qualia because being in the workspace made them available to the qualic system, whatever that might be. I think some people, those who tend to want to reduce qualia to flags or badges that give inputs a special weight, might find this kind of perspective congenial, but it doesn't appeal all that much to me. I would prefer an argument that related the appearance of qualia to a sensory input's being available to a global collection of sensory and other systems; something to do with resonances across modalities; but I happily confess I have no clear idea of how or exactly why that would work either.