Conscious Entities

The Archive

2004 to 2006

Welcome to Conscious Entities, the Archive

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 A few posts have been omitted, and quite a few are rather out of date, but I still enjoy reading them.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

Before getting on to the regular blog posts, I have included a selection of the introductory pieces about aspects of consciousness and the theories that describe them which mostly appeared as free-standing pages.

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.

One possible source of confusion is that some of the discussions here, especially in the early days ones, are presented as dialogues between two different characters. One of these, whom I think of as 'Blandula’, after the Emperor Hadrian’s verse addressed to his own soul, is represented by a sort of cherub, and is suspicious of reductive and materialist ideas: the other, 'Bitbucket’, (represented by an abacus) takes the opposite view. I hoped this would help both sides to benefit from vigorous advocacy, but before very long I stopped all that and just spoke for myself.

I hope, in any case, that you find something here interesting or amusing. As with the blog, I would ask readers to kindly make allowance for my amateurism, restricted intelligence, incompetence, sloth, and generally bad attitude to work and customer service.

This is the poem carved on Hadrian’s tomb.

Animula vagula blandula,
Hospes comesque corporis,
Quae nunc abibis in loca,
Pallidula, rigida, nudula,
Nec, ut soles, dabis iocos.

Hundreds of different translations of this poignant verse have been offered, but that isn’t going to stop me putting out my own - a fairly free version…

My little pale and wandering soul
My body’s guest and friend,
Wherever will you go to now,
When we have reached the end?
So white and clenched now you’re outside,
No covering for you;

No body now to share some fun,
The way we used to do

Definitions of consciousness

1 January 2004

Consciousness is notoriously difficult to define, though as some have pointed out, we all know what it is from direct experience. Here is a selection of notable views.

"A mere echo, the faint rumour left behind by the disappearing 'soul' upon the air of philosophy ."

William James, 'Does Consciousness Exist?'' 1904

"Consciousness: the having of perceptions, thoughts and feelings; awareness. The term is impossible to define except in terms that are unintelligible without a grasp of what consciousness means... Nothing worth reading has been written about it."

Stuart Sutherland, 'Consciousness', International Dictionary of Psychology', 1995

"...there is no question of there being a commonly accepted, exclusive sense of the term... I prefer to use it as synonymous with 'mental phenomenon' or 'mental act.' "

"...intentional in-existence, the reference to something as an object, is a distinguishing characteristic of all mental phenomena. No physical phenomenon exhibits anything similar."

Franz Brentano, 'Psychology from an empirical standpoint', 1874

" ...current values of parameters governing the high-level computations of the operating system."

P. Johnson-Laird, 'A computational analysis of consciousness', Cognition and Brain Theory, 1983

"Consciousness is like the Trinity; if it is explained so that you understand it, it hasn't been explained correctly ."

R.J. Joynt, 'Are Two Heads better than One?', Behavioural Brain Sciences, 1981

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

John Searle , 'The Mystery of Consciousness', 1997

"What is meant by consciousness we need not discuss - it is beyond all doubt."

Sigmund Freud, 'New Introductory Lectures on Psychoanalysis', 1933

"It is often held therefore (1) that a mind cannot help being constantly aware of all the supposed occupants of its own private stage, and (2) that it can also deliberately scrutinize by a species of non-sensuous perception at least some of its own states and operations. Moreover, both this constant awareness (generally called 'consciousness'), and this non-sensuous inner perception (generally called 'introspection') have been supposed to be exempt from error.

Gilbert Ryle, 'The Concept of Mind', 1949

"...perhaps 'consciousness' is best seen as a sort of dummy term like 'thing'; useful for the flexibility that is assured by its lack of specific content."

Kathleen Wilkes, 'Is Consciousness Important?', British Journal of Philosophy of Science, 1984

"We are conscious of something, on this model, when we have a thought about it. So a mental state will be conscious if it is accompanied by a thought about that state...The core of the theory, then is that a mental state is a conscious state when, and only when, it is accompanied by a suitable HOT [Higher Order Thought]"

David M. Rosenthal, 'A Theory of Consciousness', The Nature of Consciousness (ed Block, Flanagan and Güzeldere), 1997

"Behaviourism claims that consciousness is neither a definite nor a usable concept. The behaviourist, who has been trained always as an experimentalist, holds, further, that belief in the existence of consciousness goes back to the ancient days of superstition and magic."

John Watson, 'Behaviourism', 1924

"'Consciousness' is a word worn smooth by a million tongues. Depending upon the figure of speech chosen it is a state of being, a substance, a process, a place, an epiphenomenon, an emergent aspect of matter, or the only true reality."

George Miller, 'Psychology: the Science of Mental Life", 1962

"Consciousness, we shall find, is reducible to relations between objects, and objects we shall find to be reducible to relations between different states of consciousness; and neither point of view is more nearly ultimate than the other."

T.S.Eliot (Doctoral dissertation), 1916

"The concept of consciousness is a hybrid or better, a mongrel concept: the word 'consciousness' connotes a number of different concepts and denotes a number of different phenomena... P-consciousness is experience...A is access-consciousness. A state is A-conscious if it is poised for free use in reasoning and for direct 'rational' control of action and speech...Conflation of P-consciousness and A-consciousness is ubiquitous in the burgeoning literature on consciousness..."

Ned Block, 'On a Confusion about a Function of Consciousness', Behavioural and Brain Sciences, 1995

"... in the most interesting sense of the word 'consciousness', consciousness is the cream on the cake of mentality, a special and sophisticated development of mentality. It is not the cake itself."

David Armstrong, 'What is consciousness?' , The nature of Mind and Other Essays, 1980

"The improvements we install in our brain when we learn our languages permit us to review, recall, rehearse, redesign our own activities, turning our brains into echo chambers of sorts, in which otherwise evanescent processes can hang around and become objects in their own right. Those that persist the longest, acquiring influence as they persist, we call our conscious thoughts."

Daniel Dennett , 'Kinds of Minds', 1996

"The presence of mental images and their use by an animal to regulate its behavior, provides a pragmatic working definition of consciousness"

D.R.Griffin, 'The Question of Animal Awareness', 1976

Strange ideas

2 January 2004

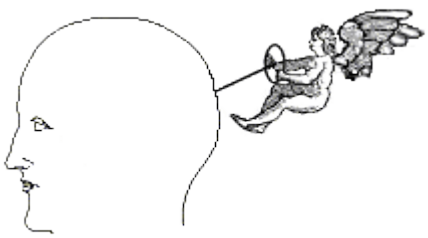
In grappling with deep philosophical issues, we are sometimes pushed towards theories we should never have adopted for their own sake. The ideas that follow are classic cases of this kind - if you find your theory leads you into one of these positions, it's a sign that you need to look again at your theory...

Homunculi

An homunculus is a 'little man'. It's evident that we don't literally have a small but fully formed person in our heads controlling what we do, but some explanations of consciousness fail because they include a central observer which itself has all the mental properties a conscious person would have, in effect an homunculus. Explaining such a central person is obviously as difficult as explaining consciousness in the first place, and the hypothesis therefore does not get us very far. The normal argument, in fact, is that homuncular arguments actually fall into an infinite regress, with the consciousness of each homunculus explained by ever smaller homunculini, nested like Russian dolls: but in fact this only applies to the hard-line homuncularist position that homunculi are the only possible explanation of consciousness. One can still believe that there is, as a matter of fact, an homunculus who is responsible for our consciousness, but that his consciousness is explained on some other basis. Most would agree, after all that there is a central entity inside us which does all the thinking - namely the brain.

Generally, however, homunculi are regarded as evidently absurd, and used mainly as a smear against other people's theories. Daniel Dennett is an unexpected exception here. No-one, in principle, is more hostile than Dennett to the idea of a central observer: Consciousness Explained repeatedly denounces the idea of the 'Cartesian Theatre', whose audience must surely be an homunculus (although the same book proposes a 'Joycean Narrator', and one might question whether the substitution of an Irish novel for a French play makes all that much difference). At the same time, Dennett has claimed that cognitive scientists frequently use homunculi in their tentative theories, and that their practice demonstrates that it is alright to do so. The apparent inconsistency is explained by the loose interpretation which Dennett gives to 'homunculus': he really means something more like a 'black box': part of a process which cannot yet be explained, but which we take for granted for the time being in order to get on with developing the rest of a theory in the meantime. Strictly, to be an homunculus, a component has to be fully conscious in its own right, or at least have a full set of conscious mental abilities, not just a particular narrowly-defined function.

Dualism



Dualism is the belief that there are two different worlds, or two fundamentally different kinds of stuff, in existence: matter and spirit, for example. In spite of being the received wisdom for hundreds of years, it is actually unsustainable in its strict form. The problem arises when you consider how the two kinds of stuff, or the two worlds, interact.

If they interact directly, they are in effect parts of a single world after all, or different forms of essentially similar stuff. If spirits affect the behaviour of material objects, they can best be thought of as another aspect of the physical world, albeit one which to date remains relatively obscure. Matter and energy, after all, are regarded as parts of the same world. It does not help to postulate intermediary stuff linking two worlds or two substances because the same problems recur for the relationship between the stuff in each world and the intermediary stuff. If, on the other hand, the two worlds don't interact, one of them is effectively null. We, the observers, must exist in one or other of the worlds (or be made of one or other kind of stuff): since the other stuff or the other world cannot affect us, we can never be aware of it and talking about it at all is a waste of time.

For reasons of this kind, and because it has sometimes been a refuge for people who would prefer things to remain unexplained, dualism has become a kind of smear which philosophers sometimes hurl at each other's theories. Most people believe the world is complex to some degree, and most accept that different levels of explanation are required in order to say all that is worth saying about it, which means there is always some hook in anyone's theory on which an accusation of dualism can be hung. There are, in any case, many different varieties of dualism, of differing degrees of radicalism. A distinction is normally drawn, in particular, between full-blown substance dualism, and property dualism, which recognises only one underlying kind of stuff, but asserts that it has two unrelated sets of properties.

In recent years, Descartes has rather unjustly come in for particularly heavy criticism over his dualism, and it seems to have become almost mandatory to start any popular book about consciousness by criticising him for it (while borrowing one of the quaint illustrations from his books). He is often baselessly accused of originating the whole idea or of having invented, rather than perpetuated, the distinction between body and soul. Of course he was a dualist, but in that respect he was merely giving a slightly more secular expression to a belief which was pretty well universal for centuries before (and after) his time. It might be fairer to see him as restricting the role of the spirit; he believed that most actions of the human body were purely mechanical, rather than directly actuated by spiritual influence, so he deserves respect from materialists.

Dualism still has its appeal: the Platonic form which gives numbers and abstract entities a real existence in a world of their own seems to appeal to mathematicians, and physicists talk about other worlds, though in both cases it is often unclear how far this is a metaphor, or a dramatic way of talking about levels of explanation rather than a genuine ontological commitment. Two of the thinkers touched on in these pages - Roger Penrose and John Eccles - advocate theories of three worlds. In spite of the hostility of most (not all) philosophers towards dualism, moreover, discussions of alternative worlds remains a popular method of argument about possibility, identity, and indeed, consciousness. Dualism also remains a basic component of much religious and paranormal theory, and hence an influential part of the intellectual background.

Epiphenomenalism

Epiphenomenalism is the view that one's actions are not caused by one's thoughts: that we are really just passive spectators under the illusion that we control our own behaviour. The idea appeals to those who want to accept that physics explains the causes of all events, including the behaviour of human beings, while still regarding consciousness as something over and above this merely mechanical process. The idea is

lent some (much-needed) plausibility by Libet's famous experiments which seem to show that decisions are effectively made before we become aware of having made them: and by many others which demonstrate people's remarkable tendency to invent reasons, and accept responsibility for, behaviour which was actually caused by a post-hypnotic suggestion, a brain malfunction, or other factors they were not consciously aware of. Strictly, of course, speech is another form of behaviour, so a rigorous epiphenomenalist would actually have to hold that these rationalisations and confabulations were also nothing to do with the person in themselves.

Philosophically, epiphenomenalism is hard to maintain because it makes the person-in-themselves so completely irrelevant. It seems better to locate personhood somewhere in the material process, or even do without it altogether. Epiphenomenalists are faced with the task of explaining how our conscious thoughts and our physical actions remain in such close harmony, which is impossible to do without raising further difficulties. Philosophically, therefore, the idea is generally regarded as untenable (though it has some appeal for David Chalmers, for example): in psychological discussions, a looser sense of the word is sometimes used, to mean a doctrine that people merely have much less conscious control over their actions than they believe.

We could illustrate the distinction between these two readings of epiphenomenalism by drawing an analogy between a human being and a large corporation. A 'corporate epiphenomenalist' would assert that the chief executive actually had no control over the organisation. On the strict philosophical reading, this would have to mean he had no influence on the organisation at all; on this view we should find ourselves forced to admit he drew no salary (because that would affect the balance sheet, however slightly), had no office, and, in fact, was effectively invisible and intangible. On the looser reading, these difficulties would not arise: the chief executive could be a substantial member of the corporate community: he could even write the Annual Report and issue convincing press releases; he just wouldn't normally be able to get any internal memos or instructions acted upon.

Solipsism

Solipsism is the belief that nothing really exists except oneself. So far as I know there are no actual solipsists (then again there might be hundreds of silent ones - solipsism rather undercuts the incentive to publish); nevertheless the idea is interesting in that it represents one extreme version of the strategy of putting the boot on the other foot. This strategy is the one adopted by those idealists who, rather than accepting the physical world and trying to explain how consciousness goes with it, begin instead with consciousness, and demand a justification of belief in the physical world. The ultimate position in this direction must surely be to deny the existence of anything outside your own mind.

Solipsism also functions as a kind of black hole which threatens to draw in other radical and sceptical viewpoints, especially epiphenomenalism. If you're an epiphenomenalist, you must believe that any other conscious entities in the world are incapable of communicating with you, or exerting any influence on the world - so what grounds can you have for believing in them? It would be a more economical hypothesis to believe that they, and the material world, are really just figments of your own imagination.

Ultimately, it's impossible to prove that solipsism is wrong; we can only point to the regularities in our experience which are most easily explained by the existence of other entities, some of them conscious. The strongest objection is that, taken seriously enough, solipsism doesn't make any difference. So everything but me is just my dream - so what? Even dreams are entities and require explanation, and to ignore them, while we are dreaming, would be as irrational as ignoring reality while awake.

Panpsychism

Panpsychists believe that everything is to some extent conscious, or looking at it another way, that conscious entities are all there really is. This neatly reverses the usual argument by taking consciousness as given and forcing opponents to explain and justify their belief in the reality of inanimate objects. Those who lean towards relativism may find this a comfortable position because it allows them to conceive of the world as consisting only of points of view.

Equally, if you are one of the people who think real, phenomenal consciousness has no place in the chain of physical cause and effect, why shouldn't everything be conscious? It wouldn't make any perceptible difference, so there's no way you can prove it isn't the case: and if everything were conscious, it might help to make the presence of consciousness in human beings seem less anomalous. David Chalmers speaks rather wistfully about the attraction of panpsychism, without quite committing himself to it. Equally, however, there can be no good reason to believe it is the case; and the usual principle, known as Occam's Razor , is not to believe in any additional entities (such as the consciousness of stones and anvils) which aren't strictly necessary.

There are a few issues to deal with, too: if everything is conscious, what counts as a thing? If this stone is conscious, is the left-hand side of it separately conscious too? Is the strange but surely valid entity consisting of my left foot and a small portion of Tower Bridge conscious?

But the ultimate problem is that panpsychism does not really revolutionise the argument the way it promises to do. The real issue facing us is to explain the difference between being conscious and not being conscious, and in another form, this remains the problem even for those who think everything is conscious. If panpsychism is true, an anvil has a form of consciousness, but one which does not influence its behaviour. My body, which after all is a physical object too, must have this same sort of ineffectual consciousness, but I also have another kind, the kind which causes me to speak, move, and act. Explaining the relationship between these two kinds of consciousness comes down to much the same thing as explaining my consciousness from a non-panpsychist materialist position.

I'm afraid this won't do at all. Far from being a 'dead end', panpsychism and panexperientialism have been developed and championed by philosophers such as Philip Goff, and as a result are enjoying a rising tide of popularity

Not So Strange ideas

3 January 2004

Levels of explanation

Reality is not simple. In biology, for example, even the most thorough-going materialist would accept that there are interesting things to be said about cells, organisms, and species which cannot be boiled down to a bald account of atoms and forces - though at the same time we never doubt for a moment that in some sense mere physics provides all the ingredients biology requires. It is just a fact about the world that it operates on these different levels - a fact nicely captured in Mary Midgeley's slogan 'one world, but a big one'. This multiplicity of levels does not normally bother us or cause us to start fretting about the reality of cells or organisms, so why can't we just see conscious thoughts and motives as another level of explanation and leave it at that?

 For one thing, we don't know yet exactly how consciousness relates to the atoms-and-forces version. In the case of cells and organisms we understand pretty well how the parts fit together and how the properties of the cell, say, arise from the nature of the atoms making it up and the relationships between them. In the case of consciousness, we just haven't got that far yet. Even when we do, I think more will still be required by way of explanation than just saying that consciousness is a matter of different levels of interpretation. We need to know how it works. What makes matters worse is the especially large and vigorous group of false theories and ideas about consciousness that have to be denounced and eliminated. We can afford to be a bit more laid back about the materialist account of cells and organisms because it isn't seriously challenged by anyone.

Anyway, this general view of reality as having a number of different levels also provides the basis for two much-quoted ideas: reduction and emergence. A reduction of a phenomenon, or a reductive explanation, accounts for higher-level properties so thoroughly in terms of lower-level entities that it is no longer strictly necessary to consider the higher level - in fact the reality of the higher level entities may be challenged. The classic example, perhaps, is temperature and mobility of molecules. Once it was understood that the two were the same, it became clear that there really wasn't anything to temperature except the motion of molecules. Another example might be the periodic table, which allowed the whole of chemistry to be reduced to physics, and indeed, some new chemistry to be discovered by extrapolation. Sometimes theories of consciousness are criticised for being too reductive - attempting to abolish consciousness as something real in its own right. On the other hand, appropriately reductive explanations are among the best you could have.

Emergence is sort of the opposite. An emergent phenomenon is a higher-level one which could not have been foreseen from the lower-level account, or at least, one which amounts to more than the simple conjunction of lower-level entities. It's often suggested that consciousness is an emergent phenomenon, though without a theory of how it emerges, that doesn't tell us much.

Higher orders

First order thoughts are straightforward thoughts - 'There's a chair.'. Second order thoughts are thoughts about thoughts - 'I knew that was a chair ...'. Quite high-order thoughts are actually

relatively common: 'I can't imagine why you thought I even cared whether he shared my taste in chairs?' - fifth order?

Anyway, an idea which recurs in several different contexts is that consciousness has something to do with higher order stuff: either our awareness of ourselves or thoughts about thoughts. Self-awareness is often considered a special, or even the only true, form of consciousness. Great significance has been attached to experiments which seem to show that while most animals treat their reflection in a mirror as another animal, some of the great apes realise that it is themselves they see (they betray this awareness by reaching up to touch a spot on their faces which they were unaware of until they looked in the mirror). Perhaps the influence of Descartes ('I think, therefore I am' has something to do with the special place which thoughts about ourselves have been given. For sceptics it is also appealing to think that the impression of self-hood might come from turning on ourselves a faculty which developed to help us respond appropriately to other organisms. The idea that containing a representation of oneself is sufficient for consciousness has been very effectively ridiculed by Roger Penrose, however, who asks scathingly whether a video camera pointing at a mirror is to be regarded as conscious.

The idea that higher order thoughts are special - that conscious thoughts are those we are aware of thinking - remains very appealing. Higher order processes certainly crop up incidentally in many theories about the mind (Searle' s idea that intentionality is characterised by the imposition of conditions of fit on conditions of fit, for example) and few would deny that they are at least a feature of consciousness. The view that they are essential to it can be traced back as far as Locke and comes in many different forms. One school of thought here is that conscious thoughts are somehow inherently about themselves as well as about their other targets; another version, expounded by David Rosenthal, maintains that thoughts are only conscious when there is a separate higher-order thought (HOT) about them. Neither view sits completely happily with the normal perception that most of our conscious thoughts are just not that complex. Many other issues also remain open - do we need thoughts about thoughts to be conscious, or will perceptions of our thoughts do? Higher-order theories are generally compatible with other theories about the nature of consciousness, and perhaps this is in the end why they haven't made more impact - no-one has succeeded in formulating them in a way which makes them controversial enough to be really interesting.

If you ask me, these ideas are in any case inherently less interesting than they seem. The point is, second order thoughts or processes are about thoughts or processes. So to have them, you have to have aboutness - intentionality. But intentionality is one of the two big issues about consciousness, and this blather about second-orders takes it for granted rather than helping to solve it. If you like, higher-order theories are pursuing the flea on the back of the bear of intentionality. If we ever catch the bear, the fleas will come along with it, but catching the fleas wouldn't mean we'd caught the bear, though while the bear's still at large, it's just as difficult.

But you can have higher-order theories which don't presuppose intentionality. Take Edelman. According to him the process starts with the recategorisation of sensory inputs: then the categories themselves are recategorised to form concepts. Concepts of concepts take us up to

second-order consciousness and self-consciousness, and concepts form the basis for language.


OK, but Edelman has the opposite problem - not actually explaining intentionality. Why are higher-order categories any more meaningful than lower-order ones? Why would a group of neurons which responds to a group of sheep be more conceptual than a group of neurons which responds to one? I don't see it.

Souls and spirits


We are ashamed

It is remarkable how poor a showing the whole idea of immaterial spirits now gets in both science and philosophy: after all, for many centuries most Europeans, including the brightest and most sophisticated thinkers, took it for granted that the explanation of consciousness lay in a spiritual realm, whereas now, as Searle has said about the existence of God, it isn't so much that everyone is a sceptic as that the question never even arises. Given that ideas from the Christian tradition are rather poorly served, it can hardly be surprising that other religious views are scarcely reflected at all in contemporary Western discussions of consciousness, in spite of the undeniable interest of many of them. In these pages we are no better than anyone else in this respect (but at least we are ashamed of ourselves).

However, the spiritual view does have one able and well-qualified scientific champion in the shape of Sir John Eccles. His credentials, and in other respects, his scientific orthodoxy, are unchallengeable, but he holds that the materialist account of the brain falls short of a full explanation and proposes a neurologically sophisticated alternative account.

Eccles' partner in developing and promulgating his theory was Sir Karl Popper, the eminent philosopher of science, and his account has a philosophical and an empirical scientific side which require separate consideration. In this respect it resembles the views of Roger Penrose , and the two theories also have in common an appeal to quantum physics and an advocacy, not just of dualism, but of a three-world system. There the resemblance ends, however.

Scientifically, Eccles and Popper argued that there is empirical evidence for a distinction between brain and mind. When the brain is stimulated with electrodes, images and memories can be evoked in the mind without shutting off the normal mental processes of perception and recollection. Surely this represents, on the one hand, images streaming up from the purely physical brain, and on the other, perceptions streaming down from a mind which can only be immaterial? Or take the example of ambiguous drawings with more than one interpretation: the brain, by itself, chooses one or other interpretation and presents the image to the mind in that light: but we can often choose to see the ambiguous image the other way, which surely represents the intervention of something else in the purely physical processing carried out by the brain. According to Eccles, the mind has access to a range of different brain processes, and by directing attention between them (like a searchlight) brings about the unity of consciousness which would otherwise be inexplicable given the diverse and complex nature of neural processes.

Specifically, Eccles suggests that spiritual psychons interact with presynaptic vesicular grids by a process analogous to the probability fields of quantum physics. A kind of spiritual cerebral

cortex interacts with the physical one undetectably but effectively at thousands of tiny sites. This allows interaction between mind and brain without violating the conservation laws of the physical world, while preserving the autonomy of the spiritual world, or 'World 2'.

World 2 is one of three worlds proposed by Eccles and Popper: World 1 is the world of physical objects; World 2 is that of states of consciousness and subjectivity (the world of qualia, presumably); World 3 is a Platonic one of objective knowledge and culture. However, the third world does not play an essential role in the theory, with the key interest being the interaction between worlds 1 and 2.

The theory remains vulnerable to the usual avenues of attack on dualism. By confining the crucial interactions to micro-sites where the minute interventions are effectively undetectable, Eccles guarantees that there can be no clear and direct physical evidence of the interventions (one might suspect that the theory is designed this way exactly to accommodate the absence of such evidence). The actual mechanism remains rather obscure - we are offered only the analogy of probability fields. But probability fields don't cause physical events; they just describe them. If the theory is a matter of probability fields, or something like them, it's surely more a matter of another level of explanation rather than another world.

We are offered little about how World 2 might work, but the psychons seem to be very similar to the neurons with which they interact. They seem to have definite physical locations, or at least something that allows them to associate separately and individually with particular physical sites. They must also have something like standard causal interactions between themselves in order to co-ordinate their handling of thousands of different micro-sites at a time. But the contents of World 2 are not meant to be another set of objects like the objects in World 1 - they are meant to be subjective states of consciousness. Are psychons states of consciousness? Either the theory is metaphysically stranger than it seems, or something is wrong here.

Ultimately, it seems that the convictions of Eccles and Popper rest on a kind of Brentano-style incredulity - it just seems obvious to them that nothing physical can possibly account for the mental. The arguments put forward about illusions and direct stimulation of the brain merely establish that consciousness can have two levels or two sets of contents, not at all that either must be non-physical: but the weakness of the argument is hidden from its proponents by their intuitive certainty that subjective consciousness is not physical.

The theory must remain unconvincing to those who don't share this intuitive certainty: but at least it does serve to show that dualist or spiritual theories need not be a matter of ignorant superstition about 'spooks'.

Memes

In his celebrated book 'The Selfish Gene' Richard Dawkins digressed to suggest that evolution might have another string to its bow besides genetics. Perhaps there was also cultural evolution. He proposed that in culture the analogue of the gene should be a 'mimeme', a unit of imitation - though Susan Blackmore says the kind of imitation involved is not in fact mimesis. In any case, Dawkins settled on the shortened form 'meme'.

Darwin would have loved it

Dawkins' original presentation of the idea had a relatively casual tone, and he acknowledges that the analogy with normal biological evolution is not perfect. His main concern seemed to be with those

cases where some error or unwanted thought succeeds in perpetuating itself in spite of being wrong or unwelcome - tunes which stick in your mind even though you don't actually like them, for example. These are clearly the interesting cases, anyway - we don't need a special explanation for the persistence of good or attractive ideas and couldn't prove in such cases that distinctively memetic processes were at work.

Memes have been taken very seriously by other authors, however, and given a much wider application. In particular, Susan Blackmore in 'The Meme Machine' has tried to straighten out some of the issues and use the basic idea to explain a wide range of mental phenomena in memetic terms. She proposes that the self is an illusion (does this make her a ghost-writer?) generated by a self-protecting complex of memes. She believes that consciousness is an illusion too, not in the sense that it doesn't exist at all, but in the sense that we falsely attribute it to a more or less homuncular central observer (though surely belief in the self does not automatically imply belief in an homunculus or Dennettian Cartesian Theatre). She believes that there is a non-memetic bedrock of consciousness in pure phenomenal experience. This leaves her with the epiphenomenalist view that our ideas don't originate from our own real consciousness, however.

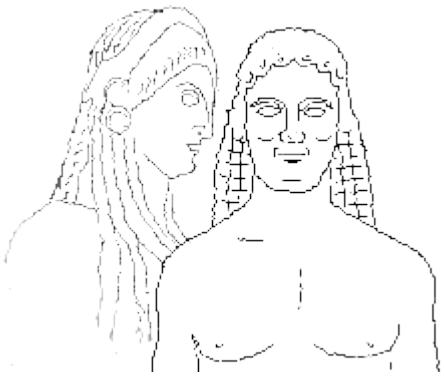
On the face of it, Dennett is bolder, asserting that our consciousness is made up of memes, or more exactly, meme effects in brains. It's a bit strange that Dennett, who already has one theory on the go, should embrace memes quite so enthusiastically. One would expect him to hold that memes are just things we attribute to each other in the course of trying to predict behaviour: that consciousness allows 'our hypotheses to die in our stead', rather then our hypotheses running us for their own benefit. The theory of the intentional stance does not seem to be dependent on memes, anyway, so perhaps it's best to assume that for Dennett memes are really just a lively metaphor...

 For Dennett, *people* are just lively metaphors...

... an illuminating way of describing processes which could, however, be fully described in other terms if necessary.

There are other objections to the whole idea of memes, but a full discussion goes beyond the scope of these pages. From our point of view the main problem is with memes as a general theory of psychology: it's too obvious that our minds don't normally work like that. There might be something interesting to be said about the evolution of urban myths, jokes, or superstitions through accumulated variations (perhaps we could tell the age of a story from the number of variations current, rather in the way biologists age hedgerows by counting the number of plan species), but it's not at all plausible to think that we conduct conversations, write novels and philosophical papers, or conduct scientific research, by copying past material over and over again until the errors accumulate into a new and better version. We do these things, not through a series of accidents, but deliberately - consciously, in fact. If memes explain anything, it must surely be unconscious processes.

Bicameralism



One of the most original theories of consciousness is the one put forward by Julian Jaynes in "The Origin of Consciousness in the Breakdown of the Bicameral Mind". It's also, at first sight, one of the least plausible: Jaynes maintains that consciousness did not arrive until some time after the writing of the Iliad. Up to that point everything done by human beings was, strictly speaking, done unconsciously.

How on earth could that be so? According to Jaynes, the minds of our ancestors were split into two distinct chambers (a bicameral structure) with one chamber in regular control and the other making occasional, authoritative interventions. What we now perceive as our own integrated conscious thoughts seemed to our more distant ancestors like voices in the head, actually from the second chamber, but attributed to spirits in the world around them; voices which they had little or no power to disobey. In later periods, the voices came to be identified as belonging to a dead king; first a recently-dead corpse propped up on a couch of stones, then an increasingly permanent god-figure in a tomb which gradually became a temple. This development coincided with, and made possible, the development of larger and more complex communities than had previously existed. Only later did people gradually come to realise that these internal voices belonged to them as much as their normal external speech.

This business of having voices in the head clearly resembles some contemporary forms of mental illness. Jaynes certainly thinks the process was subject to occasional reversion, with Joan of Arc, for example, hearing divine voices in pretty much the way her remote ancestors had done. He discusses schizophrenia and hypnosis, both possible relics of the bicameral mind in his view. He also suggests that this kind of internal conversation represents one hemisphere of the brain talking to the other.

Jaynes is a persuasive writer, with an impressive breadth of reference and a clear, rational style. It is impossible, I think, to read his book without feeling that he does have a valid point to make about ancient ways of thinking. It is a testament to his plausibility that there is a society dedicated to his views.

Nevertheless, after the immediate spell has worn off, I think his views about consciousness end up seeming almost as implausible as they did to begin with. Probably it is true that many ancient societies tended to see human behaviour as god-driven in a way which made the constant divine intervention of the Iliad a natural part of a story. By the time Virgil was writing in a similar vein, it had come to be a rather artificial literary convention: and by Swift's day it could only really be used for comic or satirical effect. But does that imply a fundamental reorganisation of the mind or just the gradual obsolescence of a literary convention? Our minds and literature are not filled with gods, but with brand names, advertising slogans, and the clichés of Hollywood and television - but would that justify a future researcher in concluding that we were unconscious robots, obeying the 'voices' of large commercial corporations? Hmm... Jaynes maintains, probably with some justice, that many ancient texts seem to display a modern cast of mind only because the translator has reinterpreted them; but surely this reinterpretation could not go so far as to make the unconscious seem conscious?

Perhaps the real sticking point is Jaynes' view that language and literature preceded consciousness. There are many tasks we can and do carry out without paying much deliberate attention, but surely speech and above all, literary composition, don't fall into that category? Far from explaining consciousness, Jaynes' view seems to require that there were two spheres of consciousness in the ancient mind. Since the god-like chamber 'spoke' to the everyday one, and the everyday one was apparently capable of writing the Iliad (and large parts of the Bible), it seems that both would have been able to pass the Turing test, anyway. When I seek an explanation for consciousness, it is, among other things, the processes of imagination and composition which go into the creation of a text which I want to have explained. Even if I agreed to stop calling these processes conscious, the essential mystery would remain untouched.

Stories about Consciousness

4 January 2004

Mary sees red

Mary the colour scientist

Once upon a time, there was a very eminent scientist called Mary, who knew everything there is to know - everything - about the physical side of colour. She understood neurology and optics, diseases of the eye, practical opthalmology, and the underlying quantum physics, all in full detail, and much more besides. Her memory and understanding were extraordinary - in fact, some people who don't believe in the story have said that they must have been so extraordinary that really, we just confuse ourselves by attributing such amazing knowledge to a human being. Be that as it may, by a strange quirk of fate, Mary had never seen colours - although her colous vision functioned perfectly well. She had spent her life confined in a room where nothing was coloured (what is it with these philosophers about locking people in rooms?): all she could ever see was black, white, and different shades of grey.

One day, however, her colleagues at the Institute of Vision finally decided it was time to let Mary out, and led her into a room where a single red rose stood in a vase on the table. Suddenly she could see colours, or at least a colour, for the first time. Now the point is this: I said that Mary knew everything about colour from the objective, physical point of view, but when she saw it for the first time, she knew something she'd never known before - what it is like to see colour. Didn't she? And this proves that really seeing red involves something over and above the simple business of wavelengths and electrical impulses. Doesn't it?

No, of course not. Mary acquired no new knowledge when she saw the rose - she had simply had a new experience. Focussing too exclusively on the role of the senses as information gatherers can lead us into the error of supposing that to experience a particular sight or sound is merely to gain some information. If that were so, reading the label on a bottle of wine would be as enjoyable as drinking it. Of course experiencing something allows us to generate information about it, but we also experience the reality , which in itself has nothing to do with information. I suspect that academics are particularly prone to this confusion. Perhaps they have spent so much time dealing with abstractions that reality begins to seem puzzling, and the fact that information about the colour red does not contain any real redness comes to seem a philosophical puzzle instead of the banal fact it really is. This is not to say that there are no mysteries about the nature of reality: but the fact that it doesn't consist of information isn't one.

There's none so blind. However, I can't quite leave it there - there's another point which really has to be made. It's fairly obvious that when Frank Jackson came up with this story, he meant to be politically correct by choosing a woman as the expert colour scientist. But what was the result? Instead of merely perpetuating the assumption that scientists are men, he ended up painting a picture of a woman routinely confined, controlled, and turned into an experimental animal - a far worse revelation of grossly unacceptable subconscious attitudes.

Bill the chip-head

Once upon a time, in the dead of night, a gang of mad, evil scientists (in philosophers' stories, scientists are usually mad and evil) crept into Bill's room, and carefully anaesthetising him, opened up his skull. With infinite care, they placed in his brain a tiny device which, when switched on remotely, would take over the work of a single neuron. Although the device was based on a silicon chip, it reproduced the functional behaviour of the neuron perfectly, so that whether Bill's brain was using the original neuron or the new chip, its behaviour would remain exactly the same. Repairing Bill's skull with such exquisite skill that no detectable trace of the operation remained, the triumphant scientists hurried away.

Bill noticed nothing

Over the course of the next few months, they returned on many occasions, replacing thousands of neurons until the unsuspecting Bill had, in effect, two brains. By throwing the lever in their secret laboratory, the scientists could switch Bill from operating with a normal brain, to operating with one composed entirely of silicon chips. Not only had they demonstrated the possibility of an artificial brain: they had produced one so similar to Bill's real brain that they could switch between the two while Bill was in the middle of a sentence without causing so much as a momentary pause.

But what was happening to BIll's qualia ? If qualia, real subjective experiences, only come from proper human brains, then every time the scientists switched to the silicon brain, they must disappear. They could waggle the switch on and off and produce 'dancing qualia' if they wanted. Not only that - how did it work when they were only half-way through the replacement programme, with only half the neurons affected? Did Bill have faded qualia, half as intense as before? Or were certain neurons crucial, making a sudden, absolute difference?

The really bizarre thing is that none of this affected Bill's behaviour in the least. With his silicon brain switched on, his behaviour was exactly the same as it would have been with it off - so although in one case he wasn't having a real experience of redness at all, he would still say that he was. The only conclusion you can draw is that there can't be anything fundamentally special about neurons after all. If qualia aren't constituted by functions, they must at least be determined by them, so that a silicon brain with the same functional patterns gives rise to just the same qualia as an organic one. Doesn't it?

That just begs the whole question. By assuming that silion chips can stand in for neurons, you build in the answer you want from the start. The trouble is, people like you have been assuming for a long time that neurons are just electric switches. The whole neural network thing is based on that assumption. But it isn't anything like as simple as that. Synapses depend on some very sophisticated chemistry, with many different transmitters and other factors to take into account. The way a neuron functions depends intimately on all sorts of messy biological factors, quite unlike a chip. I don't think it's even safe to ignore the role of glial cells, which are usually just dismissed as packaging for the neurons, but actually play an important role in the biochemical environment of the brain. It really is strange the way people attempt to explain things like emotions in terms of wiring, when it's so evident that hormones and other chemical factors have a leading role.

 You're misreading the story. Two things. First, of course this kind of experiment isn't remotely practical, and never will be. There are all sorts of difficulties. We could work on the story a bit to remove some of the plausibility problems - Bill could have been paralysed in an accident and kept in hospital permanently, for example. But it doesn't matter really - in these examples we have to be allowed a bit of magic. The point is, what would be the consequences if this were possible? If you refuse to consider anything but true examples, we're not going to get very far.

Second, I don't assume that neurons are just electric switches. This chip device could have tiny reservoirs of all the necessary neurotransmitters, or whatever. You just have to accept, as part of the starting assumptions, that in all the relevant respects, it performs the functions of a neuron. If it helps, you can suppose that the chips replace groups of neurons, rather than single ones.

 No, it won't work! I realise we have to accept some unrealistic premises for the sake of argument at times, but this is different. You explicitly took it for granted just then that there is some set of properties of Bill's neurons that include all the relevant properties, and that the rest are unimportant. But to assume that is to assume functionalism or something like it from the very start! All the properties of Bill's brain are potentially relevant - none of them can be excluded. A robot, a machine, or a computer, are what they are because they approximate to some abstract design: but Bill is just that organism over there. He's not an approximate instance of Billhood, he's Bill. That's exactly what makes him a real person, with real experiences, instead of a simulation. You must see that, surely?

 I see that this is some weird kind of essentialism. Just look at the human body - even you won't deny that it has organs and structures which are for particular jobs. They may not literally have been designed, but in many ways they look as if they were. It's the same with the brain. We don't know all the answers yet, but surely it's clear enough that certain parts of the brain do certain jobs, and that neurons were meant, in some sense, to work a particular way. What if Bill spontaneously develops a cancer in his brain which makes his speech garbled. Are you going to say that's all part of his nature, and perfectly OK? It seems to me that in such cases, something has clearly gone wrong with the brain, in the same way things can go wrong with a designed machine - and sometimes we can even put that kind of thing right again.

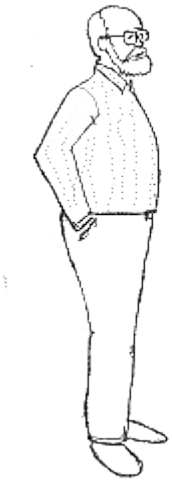
Anyway, in a sense, he is an approximate instance of Billhood. As a phenotype, he is one instantiation of the design embodied in his genotype - his DNA.

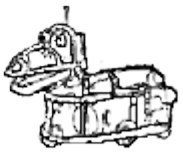
 No way. If that were true, Bill and his identical twin would share a single identity...

Key Theories and Theorists

Daniel Dennett - the intentional stance

5 January 2004





Dennett is the great demystifier of consciousness. According to him there is, in the final analysis, nothing fundamentally inexplicable about the way we attribute intentions and conscious feelings to people. We often attribute feelings or intentions metaphorically to non-human things, after all. We might say our car is a bit tired today, or that our pot plant is thirsty. At the end of the day, our attitude to other human beings is just a version - a much more sophisticated version - of the same strategy. Attributing intentions to human animals makes it much easier to work out what their behaviour is likely to be. It pays us, in short, to adopt the intentional stance when trying to understand human beings.

This isn't the only example of such a stance, of course. A slightly simpler example is the special 'design stance' we adopt towards machines when we try to understand how they work (that is, by assuming that they do something useful which can be guessed from their design and construction). An axe is just a lump of wood and iron, but we naturally ask ourselves what it could be for, and the answer (chopping) is evident. A third stance is the basic physical one we adopt when we try to predict how something will behave just by regarding it as a physical object and applying the laws of physics to it.

It's instructive to notice that when we adopt the design stance towards an axe, we don't assume that the axe is magically imbued with spiritual axehood: but at the same time its axehood is uncontroversially a fact. If we only understood things this way all the time, we should find the real nature of people and thoughts no more worrying than the real nature of axes.

One day there could well be machines which fully justify our adopting the intentional stance towards them and hence treating them like human beings. With some machines, some of the time, and up to a point, we do this already, (think of computer chess) but Dennett would not predict the arrival of a robot with full human-style consciousness for a while yet.

So it's all a matter of explanatory stances. But doesn't that mean that people are not 'real', just imaginary constructions? Well, are centres of gravity real? We know that forces really act on every part of a given body, but it makes it much easier, and no less accurate, if our calculations focus on a single average point. People are a bit like that. There are a whole range of separate processes going on in the relevant areas of your brain at any one time - producing a lot of competing 'multiple drafts' of what you might think, or say. Your actual thoughts or speech emerge from this competition between rival versions - a kind of survival of the fittest, if you like. The intentional stance helps us work out what the overall result will be.

 The 'overall result'? But it's not as if the different versions get averaged out, is it? I thought with the multiple drafts idea one draft always won at the expense of all the others. That's one of the weaknesses of the idea - if one 'agent' can do the drafting on its own, why would you have several?

 It's just more effective to have several competing drafts on the go, and then pick the best. It's a selective process, comparable in some respects to evolution - or a form of parallel processing, if you like.

 'Pick the best'? I don't see how it can be the best in the sense of being the most cogent or useful thought or utterance - it's just the one that grabs control. The only way you could guarantee it was the best would be to have some function judging the candidates. But that would be the kind of central control which the theory of multiple drafts is supposed to do away with. Moreover, if there is a way of judging good results, there surely ought to be a way of generating only good ones to begin with - hence again no need for the wasteful multiple process. I'm always suspicious when somebody invokes 'parallel processing' . At the end of the day, I think you're forced to assume some kind of unified controlling process.

 Absolutely not- and this is a key point of Dennett's theory. None of this means there's a fixed point in the brain where the drafts are adjudicated and the thinking gets done. One of the most seductive delusions about consciousness is that somewhere there is a place where a picture of the world is displayed for a 'control centre' to deal with - the myth of the 'Cartesian Theatre'. There is no such privileged place; no magic homunculus who turns inputs into outputs. I realise that thinking in terms of a control centre is a habit it's hard to break, but it's an error you have to put aside if we're ever going to get anywhere with consciousness.

Another pervasive error, while we're on the subject, is the doctrine of 'qualia' - the private, incommunicable redness of red or indescribable taste of a particular wine. Qualia are meant to be the part of an experience which is left over if you subtract all the objective bits. When you look at something blue, for example, you acquire the information that it is blue: but you also, say the qualophiles, see blue. That blue you really see is an example of qualia, and who knows, they ask, whether the blue qualia you personally experience are the same as those which impinge on someone else?

Now qualia cannot have any causal effects (otherwise we should be able to find objective ways of signalling to each other which quale we meant). This has the absurd consequence that any words written or spoken about them were not, in fact, caused by the qualia themselves. There has been a long and wearisome series of philosophical papers about inverted spectra, zombies, hypothetical twin worlds and the like which purport to prove the existence of qualia. For many people, this first person, subjective, qualia-ridden experience is what consciousness is all about; the mysterious reason why computers can never deserve to be regarded as conscious. But, Dennett says, let's be clear: there are no such things as qualia. There's nothing in the process of perception which is ultimately mysterious or outside the normal causal system. When I stand in front of a display of apples, every last little scintilla of subtle redness is capable of influencing my choice of which one to pick up.


It's easy to deny qualia if you want to. In effect you just refuse to talk about them. But it's a bit sad. Qualia are the really interesting, essential part of consciousness: the bit that really matters. Dennett says we'll be alright if we stick to the third-person point of view (talking about how other people's minds work, rather than talking about our own); but it's our own, first-person sensations and experiences that hold the real mystery, and it's a shame that Dennett should deny himself the challenge of working on them.


I grant you qualia are grist to the mill of academic philosophers - but that's never been any sign that an issue was actually real, valid, or even interesting. But in any case, Dennett hasn't excluded himself from anything. He proposes that instead of mystifying ourselves with phenomenology we adopt a third-person version - heterophenomenology. In other words, instead of trying to talk about our ineffable inner experiences, we should talk about what people report as being their ineffable inner experiences. When you think about it, this is really all we can do in any case.

That's Dennett in a nutshell. Actually, it isn't possible to summarise him that compactly: one of his great virtues is his wide range. He covers more aspects of these problems than most and manages to say interesting things about all of them. Take the frame problem - the difficulty computer programs have in dealing with teeming reality and the 'combinatorial explosion' which results. This is a strong argument against Dennett's computation-friendly views: yet the best philosophical exposition of the problem is actually by Dennett himself.


Mm. If you ask me, he's a bit too eager to cover lots of different ideas. In 'Consciousness Explained' he can't resist bringing in memes as well as the intentional stance, though it's far from clear to me that the two are compatible. Surely one theory at a time is enough, isn't it? Even Putnam disavows his old theory when he adopts a new one.


It seems to me that a complete account of consciousness is going to need more than one theoretical insight. Denett's broad range means he's said useful things on a broader range of topics than anyone else. Even if you don't agree with him, you must admit that that sceptical view about qualia, for example, desperately needed articulating. And it typifies the other thing I like about Dennett. He's readable, clear, and original, but above all he really seems as if he wants to know the truth, whereas most of the philosophers seem to enjoy elaborating the discussion far more than they enjoy resolving it. His theory may seem strange at first, but after a while I think it starts to seem like common sense. Take the analogy with centres of gravity. People must be something like this in the final analysis, mustn't they? On the one hand we're told the self is a mysterious spiritual entity which will always be beyond our understanding: on the other side, some people tell us paradoxically that the self is an illusion. I don't think either of these positions is easy to believe: by contrast, the idea of the self as a centre of narrative gravity just seems so sensible, once you've got used to it.


The problem is, it's blindingly obvious that whether something is conscious or not doesn't depend on our stance towards it. Dennett realises, of course, that we can't make a bookshelf conscious just by giving it a funny look, but the required theory of what makes

something a suitable target for the stance (which is really the whole point) never gets satisfactorily resolved in my view, in spite of some talk about 'optimality'.

And that business about centres of gravity. A centre of gravity acts as a kind of average for forces which actually act on millions of different points. Well there really are people like that - legal 'persons', the contractual entitities who provide a vehicle for the corporate will of partnerships, companies, groups of hundreds of shareholders and the like. But surely it's obvious that these legal fictions, which we can create or dispell arbitrarily whenever we like, are entirely different to the real people who invented them, and on whom, of course, they absolutely depend.

The fact is, Dennett's view remains covertly dependent on the very same intuitive understanding of consciousness it's meant to have superseded. You can imagine a disciple running into problems like this...

> Disciple: Dan, I've absorbed and internalised your theory and at last I really understand and believe it fully. But recently I've been having a difficulty.
>
> Dennett: What's that?
>
> Disciple: Well, I can't seem to adopt the intentional stance anymore.
>
> Dennett: Wow. It's really very simple. Deep breaths now. Look at the target (use me if you like). Now just attribute to me some plausible conscious states and intentions.
>
> Disciple: But... What would that be like? What are conscious states? For you to have conscious states just means I can usefully deal with you as if you had ... conscious states. I seem to be caught in a kind of vicious circle unless I just somehow know what conscious states are...
>
> Dennett: Steady now. Just think, what would I be likely to do if I had the kind of real, original intentions which people talk about? How would things with intentions behave?
>
> Disciple: I have no idea. There are no things with real intentions. I'm not even sure any more what 'real intentions' means...


Yes, very amusing I'm sure. I suppose I can sympathise with you to some extent. Grasping Dennett's ideas involves giving up a lot of cherished and ingrained notions, and I'm afraid you're just not ready (or perhaps able) to make the effort. But the suggestion that Dennett doesn't tell us what makes something a good target for the intentional stance is a shocking misrepresentation. It could hardly be more explicit. Anything which implements a 'Joycean machine' is conscious. This Joycean machine is the thing, the program if you like, which produces the multiple drafts. The idea is that consciousness arises when we turn on ourselves the mechanisms and processes we use to recognise and understand other people. Crudely put, consciousness is a process of talking to ourselves about ourselves: and it's that that makes us susceptible to explanation through the intentional stance. It's all perfectly clear.

You obviously haven't grasped the point about optimality, either. Suppose you're playing chess. How do you guess what the other player is likely to do? The only safe thing to do is to assume he will make the best possible move, the optimal move. In effect, you attribute to him the desire to win and the intention of out-playing you, and that helps dramatically in the task of deciding

which pieces he is likely to move. Intentional systems, entities which display this kind of complex optimality, deserve to be regarded as conscious to that extent.

 Yes, yes, I understand. But how do you know what behaviour is optimal? Things can't just be inherently optimal: they're only optimal in the light of a given desire or plan. In the case of a game of chess, we take it for granted that someone just wants to win (though it ain't necessarily so): but in real-life contexts it's much more difficult. Attributing desires and beliefs to people arbitrarily won't help us predict their behaviour. Our ability to get the right ones depends on an in-built understanding of consciousness which Dennett does not explain. In fact it springs from empathy: we imagine the beliefs and desires we would have in their place. If we hadn't got real beliefs and desires ourselves, the whole stance business wouldn't work

 It isn't empathy we rely on - at least, not what you mean by empathy. The process of evolution has fitted out human beings with similar basic sets of desires (primarily, to survive and reproduce) which can be taken for granted and used as the basis for deductions about behaviour. I don't by any means suggest the process is simple or foolproof (predicting human behaviour is often virtually impossible) just that treating people as having conscious desires and beliefs is a good predictive strategy. As a matter of fact, even attributing incorrect desires and beliefs would help us falsify some hypotheses more efficiently than trying to predict behaviour from brute physical calculation.

Speaking of evolution, it occurs to me that a wider perspective might help you see the point. Dennett's views can be seen as carrying on a long-term project of which the theory of evolution formed an important part. This is the gradual elimination of teleology from science. In primitive science, almost everything was explained by attributing consciousness or purpose to things: the sun rose because it wanted to, plants grew in order to provide shade and food, and so on. Gradually these explanations have been replaced by better, more mechanical ones. Evolution was a huge step forward in this process, since it meant we could explain how animals had developed without the need to assume that conscious design was part of the process. Dennett's work takes that kind of thinking into the mind itself.

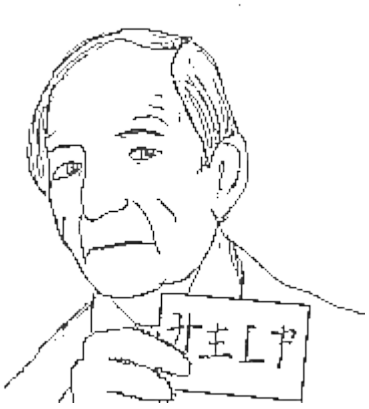
 Yes, but absurdly! It was fine to eliminate conscious purposes from places where they had no business, but to eliminate them from the one place where they certainly do exist, the mind, is perverse. It's as though someone were to say, well, you know, we used to believe the planets moved because they were gods; then we came to realise they weren't themselves conscious beings, but we still believed they were moved by angels. After a while, we learnt how to do without the angels: now it's time to take the final step and admit that, actually, the planets don't move. That would be no more absurd that Dennett's view that, as he put it, 'we are all zombies'.

 A palpably false analogy: and as for the remark about zombies, it is an act of desperate intellectual dishonesty to quote that assertion out of context!

John Searle - the Chinese Room

6 January 2004


Searle is a kind of Horatius, holding the bridge against the computationalist advance. He deserves a large share of the credit for halting, or at least checking, the Artificial Intelligence bandwagon which, until his paper 'Minds, Brains and Programs' of 1980 seemed to be sweeping ahead without resistance. Of course, the project of "strong AI" (a label Searle invented), which aims to achieve real consciousness in a machine, was never going to succeed, but there has always been (and still is) a danger that some half-way convincing imitation would be lashed together and then hailed as conscious. The AI fraternity has a habit of redefining difficult words in order to make things easier. Terms for things which, properly understood, imply understanding, and which computers can't, therefore, handle - are redefined as simpler things which computers can cope with. At the time Searle wrote his paper, it looked as if "understanding" might quickly go the same way, with claims that computers running certain script-based programs could properly be said to exhibit at least a limited understanding of the things and events described in their pre-programmed scenarios. If this creeping debasement of the language had been allowed to proceed unchallenged, it would not have been long before 'conscious', 'person' and all of the related moral vocabulary were similarly subverted, with dreadful consequences.



After all, if machines can be people, people can be regarded as merely machines, with all that implies for our attitude to using them and switching them on or off


Are you actually going to tell us anything about Searle's views, or is this just a general sermon?


Searle's main counter-stroke against the trend was the famous 'Chinese Room' . This has become the most famous argument in contemporary philosophy; about the only one which people who aren't interested in philosophy might have heard of. A man is locked up, given a lot of data in Chinese characters, and runs by hand a program which answers questions in Chinese. He can do that easily enough (given time), but since he doesn't understand Chinese, he doesn't understand the questions or the answers he's generating. Since he's doing exactly what a computer would do, the computer can't understand either.


The trouble with the so-called Chinese Room argument is that it isn't an argument at all. It's perfectly open to us to say that the man in the machine understands the Chinese inputs if we want to. There is a perfectly good sense in which a man with a code book understands messages in code.

However, that isn't the line I take myself. It's clear to me that the 'systems' response, which Searle quotes himself, is the correct diagnosis. The man alone may not understand, but the

man plus the program forms a system which does. Now elsewhere, Searle stresses the importance of the first-person point of view, but if we apply that here we find he's hoist with his own petard. What's the first-person view of whatever entity is answering the questions put to the room? Suppose instead of just asking about the story, we could ask the room about itself: who are you, what can you see? Do you think the answer would be 'I'm this man trapped in a room manipulating meaningless symbols'? Of course not. To answer questions about the man's point of view, the program would need to elicit his views in a form he understood, and if it did that it would no longer be plausible that the man didn't know what was going on. The answers are clearly coming from the system, or in any case from some other entity, not from the man. So it isn't the man's understanding which is the issue. Of course, the man, without the program, doesn't understand. In just the same way, nobody claims an unprogrammed computer can understand anything.

But even as a purely persuasive story, I don't think it works. Searle doesn't specify how the instructions used by the man in the room work: we just know they do work. But this is important. If the program is simple or random, we probably wouldn't think any understanding was involved. But if the instructions have a high degree of complexity and appear to be governed by some sophisticated overall principle, we might have a different view. With the details Searle gives, I actually think it's hard to have any strong intuitions one way or the other.


Actually, Searle never claimed it was a logical argument, only a gedankenexperiment. So far as details of how the instructions work, it's pretty clear in the original version that Searle means the kind of program developed by Roger Schank: but it doesn't matter much, because it's equally clear that Searle draws the conclusion for any possible computer program.

Whatever you think about the story's persuasiveness, it has in practice been hugely influential. Whether they like it or not (and some of them certainly don't), all the people in the field of Artificial Intelligence have had to confront it and provide some kind of answer. This in itself represented a radical change; up to that point they had not even had to talk about the sceptical case. The angriness of some of the exchanges on this subject is remarkable (it's fair to say that Searle's tone in the first place was not exactly emollient) and Searle and Dennett have become the Holmes and Moriarty of the field - which is which depends on your own opinion. At the same time, it's fair to say that those of a sceptical turn of mind often speak warmly of Searle, even if they don't precisely agree with him - Edelman, for example, and Colin McGinn . But if the Chinese Room specifically doesn't work for you, it doesn't matter that much. In the end, Searle's point comes down to the contention - surely unarguable - that you can't get syntax from semantics. Just shuffling symbols around according to formal instructions can never result in any kind of understanding.


But that is what the whole argument is about! By merely asserting that, you beg the question. If the brain is a machine, it seems obvious to me that mechanical operations must be capable of yielding whatever the brain can yield.


Well, let's try a different tack. The Chinese Room is so famous, it tends to overshadow Searle's other views, but as you mentioned, he puts great emphasis on the first-person perspective and regards the problem of qualia as fundamental. In fact, in arguing with Dennett,

he has said that it is the problem of consciousness. This is perhaps surprising at first glance, because the Chinese Room and its associated arguments about semantics are clearly to do with meaning, not qualia. But Searle thinks the two are linked. Searle has detailed theories about meaning and intentionality which are arguably far more interesting (and if true, important) than the Chinese Room. It's difficult to do them justice briefly, but if I understand correctly, he analyses meaning in terms of intentionality (which in philosophy means aboutness), and intentionality is grounded in consciousness. How the consciousness gets added to the picture remains an acknowledged mystery, and actually it's one of Searle's virtues that he is quite clear about that. His hunch is that it has something to do with particular biological qualities of the brain, and he sees more scientific research as the way forward.

One of Searle's main interests is the way certain real and important entities (money, football) exist because someone formally declared that they did, or because we share a common agreement that they do. He thinks meaning is partly like that. The difference between uttering a string of noises and meaning something by them is that in the latter case we perform a kind of implicit declaration in respect of them. In Searle's terminology, each formula has conditions of satisfaction, the conditions which make it true or false: when we mean it, we add conditions of satisfaction to the conditions of satisfaction. This may sound a bit obscure, but for our purposes Searle's own terminology is dispensable: the point is that meaning comes from intentions. This is intuitively clear - all it comes down to is that when we mean what we say, we intend to say it.

So where does intentionality, and intentions in particular, come from? The mystery of intentionality - how anything comes to be about anything - is one of the fundamental puzzles of philosophy. Searle stresses the distinction between original and derived intentionality. Derived intentionality is the aboutness of words or pictures - they are about something just because someone meant them to be about something or interpreted them as being about something: they get their intentionality from what we think about them. Our thoughts themselves, however, don't depend on any convention, they just are inherently about things. According to Searle, this original intentionality develops out of things like hunger. The basic biochemical processes of the brain somehow give rise to a feeling of hunger, and a feeling of hunger is inherently about food.

Thus, in Searle's theory, the two basic problems of qualia and meaning are linked. The reason computers can't do semantics is because semantics is about meaning; meaning derives from original intentionality, and original intentionality derives from feelings - qualia - and computers don't have any qualia. You may not agree, but this is surely a most comprehensive and plausible theory.


Except that both qualia and intrinsic intentionality are incoherent myths! How can anything just be inherently about anything? Searle's account falls apart at several stages. He acknowledges he has no idea how the biomechanical processes of the brain give rise to 'real feelings' of hunger, and he also has no account of how these real feelings then prompt action. In fact, of course, the biomechanical story of hunger does not suddenly stop at some point: it flows on smoothly into the biomechanical processes of action, of seeking food and of eating. Nothing in that process is fundamentally mysterious, and if we want to say that a real feeling of hunger is involved in causing us to eat, we must say that it is part of that fully-mechanical, computable, non-mysterious process - otherwise we will be driven into epiphenomenalism.

When you come right down to it, I just do not understand what motivates Searle's refusal to accept common sense. He agrees that the brain is a machine, he agrees that the answer is

ultimately to be found in normal biological processes, and he has a well-developed theory of how social processes can give rise to real and important entities. Why doesn't he accept that the mind is a product of just those physical and social processes? Why do we need to postulate inherent meaningfulness that doesn't do any work, and qualia that have no explanation? Why not accept the facts - it's the system that does the answering in the Chinese Room, and it's a system that does the answering in our heads!

 It is not easy for me to imagine how someone who was not in the grip of an ideology would find that idea at all plausible!

David Chalmers - the Hard Problem

7 January 2004

With 'The Conscious Mind: In Search of a Fundamental Theory' David Chalmers introduced a radical new element into the debate about consciousness when it was perhaps in danger of subsiding into unproductive trench warfare. Many found some force in his arguments; others have questioned whether they are particularly new or effective, but even if you don't agree with him, the energising effect of his intervention can still be welcomed. Chalmers believes (and of course he's not alone in this respect) that there are two problems of consciousness. One is to do with how sensory inputs get processed and turned into appropriate action; the other is the problem of qualia - why is all that processing accompanied by sensations, and what are these vivid sensations, anyway? He calls the first the 'easy' problem and the second, which is the real focus of his attention, the 'hard' problem. Chalmers is careful to explain that he doesn't mean the 'easy' problem is trivial, just nothing like as mind-boggling as qualia, the redness of red, the ineffably subjective aspect of experience.

The real point, in any case, is his view of the 'hard' problem, and here the unusual thing about Chalmers' theory is the extent to which he wants to take on two views which are normally seen as opposed. He wants behaviour to be explainable in terms of a materialist, functionalist theory, operating within the normal laws of physics: in fact, he ends up seeing no particular barrier to the successful creation of consciousness in a computer. But he also wants qualia which remain mysterious in some respects and which appear to be have no causal effects. He doesn't quite commit himself on this last point: the causal question remains open (qualia might over-determine events, for example, having a causal influence which is always in the shadow of similar influences from straightforward physical causes) and he does not sign up explicitly to epiphenomenalism (the view that our thoughts actually have no influence on our actions) - but he thinks the current arguments for the opposite views are faulty. All the words in the mental vocabulary, on his view, acquire two senses: there is psychological pain, for example, which plays a full normal part in the chain of cause and effect, and affects our behaviour: and then there is phenomenal pain, which does not determine our actions, but which actually, you know, hurts.

Chalmers is surely a dualist, because he believes in two kinds of fundamental stuff, and he is an epiphenomenalist, because he believes our thoughts and feelings have no real influence on the world. Neither of these positions makes sense. The book pulls its punches in these kinds of areas. He says he does not describe his view as epiphenomenalism, but that the alternatives to epiphenomenalism are wrong. Now if you believe the negation of a view is wrong, you have to believe the view is right, don't you? And what is this 'causal over-determination' business? So an event is caused by some physical prior event, and also caused by the qualia - but it would have happened just the same way if the qualia weren't there? Chalmers says there's

no proof this is true, but no real argument to disprove it, either. How about Occam's Razor? A causal force which makes no difference to events is a redundant entity which ought to be excised from the theory. Otherwise we might as well add undetectable angels to the theory - hey, you can't prove they don't exist, because they wouldn't make any difference to anything anyway.

 This aggressive attitude is out of place. I think you have to take on board that Chalmers is quite honest about not presenting a final answer to everything. What he's about is taking the problems seriously. This has a certain resonance with many people. There was a gung-ho era of artificial intelligence when many people just ignored the philosophical problems, but by the time Chalmers published "The Conscious Mind" I think more were prepared to admit that maybe the problem of qualia was more substantial than they thought. Chalmers seemed to be speaking their language. Of course, this may be irritating to philosophers who may feel they had been going on about qualia for years without getting much attention. It irritates some of the philosophers even more (not necessarily a bad thing) when Chalmers adopts (or fails definitely to reject, anyway) views like epiphenomenalism, which they mostly regard as naive. But you really can't say Chalmers is philosophically naive - he has an impressive command of technical philosophical issues and handles them with great aplomb.

 Oh, yes. All those pages of stuff about supervenience, for example. That's exactly what I hate about philosophy - the gratuitous elaboration of pointless technical issues. I mean, even if we got all that stuff straight, it wouldn't help one iota. We could spend years discussing whether, say, the driving of a car down the road supervenes under the laws of physics on the spark in the cylinder at time t, or under some conjunction of laws of modal counterfactuals, yet to be specified, with second-order laws of pragmatic engineering theory. Or some load of old tripe like that. It wouldn't tell us how the engine works - but that's what we want to know, and the same goes for the mind.

 Well, I'm sorry but you have to be prepared to take on some new and slightly demanding concepts if we're going to get anywhere. We can't get very far with naive ideas of cause and effect: the notion of supervenience gives us a way to unravel the issues and tackle them separately. I know this is difficult stuff to get to grips with, but we're talking about difficult issues here. You just want the answer to be easy.

 Easy! It's Chalmers who ignores the real problems. Look at dualism. It's only worth accepting a second kind of stuff if it makes things easier to explain. If we could solve the problem of qualia by assuming they live in a different world, there might be some point. But we can't: they're just as hard to explain in a dualist world as they were in a monist, materialist one, and on top of that you have to explain how the two worlds relate to each other. Chalmers ends up with 'bridging principles', which specify that phenomenal states always correspond with psychological ones. This sounds like Leibniz's pre-established harmony between the spirit and body, but at least Leibniz had God to arrange things for him! Chalmers actually has no way of knowing whether psychological and phenomenal states correspond, because he only ever experiences one of them (which one depends on whether it's Phenomenal Chalmers or Psychological Chalmers we're talking about, I suppose). The final irony is that it's Psychological

Chalmers who writes the books, because that's a physical, cause-and-effect matter: but his reasons for writing about qualia can't be anything to do with qualia themselves, because he never experiences them - only Phenomenal Chalmers does that... And we haven't even touched on the stuff about how thermostats feel, and the mysterious appeal of panpsychism. But really, the worst of it is that the problem he's inviting people to 'take seriously' is the wrong one. The whole 'problem of qualia' is a delusion.


On the contrary, it's the whole point. You should read less Dennett and more by other people. Incidentally, it must be in Chalmers' favour that neither Dennett nor his arch-enemy Searle has any time at all for Chalmers. He must be doing something right to attract opposition like that from both extremes, don't you think?

Two points, though. First, if we want to make any progress at all, it's going to involve contemplating some weird-looking ideas. All the mainstream ones have been done already. Chalmers is all about opening up possibilities, not presenting a cast-iron finished theory. Second, you're talking as if Chalmers took up dualism for no reason, but in fact he gives a whole series of arguments which explain why we're forced to that conclusion.

- Argument 1: The logical possibility of zombies, people exactly like us but with no qualia. This is the main one, which puts in its simplest form Chalmers' underlying point of view that qualia are separable from the normal physical account of the world, and so just must be something different.
- Argument 2: The Inverted Spectrum. An old classic, which relies on the same basic insight as the first argument, ie that you could change the qualia without changing anything else. Arguments along these lines have been elaborated to the nth degree elsewhere, but Chalmers' version is pretty clear.
- Argument 3: From epistemological asymmetry. Qualia just don't look the same from the inside. When we examine the biology of our leg, it isn't essentially different from examining someone else's: but when we examine our own sensations, it bears no resemblance to observing the sensations of others.
- Argument 4: The knowledge argument. Our old friend Mary the colour scientist .
- Argument 5: The absence of analysis. This is simply a matter of putting the onus on the opposition to give an account of how qualia could possibly be physical.

The main point of the main argument, very briefly, is that we can easily imagine a 'zombie': a person who has all the psychological stuff going on, but no subjective experience. At the very least, it's logically possible that there should be such people. As a result, you cannot just identify the physical workings of the brain, the psychological aspect, with the subjective experience, the phenomenal aspect. I have to say I think this is essentially correct.


There's no way we can know whether something is logically possible unless we understand what we're talking about. We need to know what phenomenal consciousness is before we can decide whether zombies without it are possible. Chalmers assumes it's obvious that phenomenal experience isn't physical, and hence it's obvious we could have zombies. But this just begs the question. I assume phenomenal experience is a physical process, so it's obvious to me that there couldn't, logically, be a person who was physically identical to me without them having my experiences. Look at it this way. If Chalmers didn't understand physics, he would probably find it easy to imagine that the molecules inside him could move around

faster without his temperature going up. But when he understands what temperature really is, he can see that it was logically impossible after all.

Chalmers is really presenting intuitions disguised as arguments - alright, he's not alone in that, but they're dodgy intuitions, too. Look at that stuff about information. According to Chalmers, anything with a shape or marks on it, in fact anything at all, is covered in information - information about itself and how it got the way it is. We can speculate that any kind of information might give rise to consciousness: maybe even thermostats have a dim phenomenal life similar to just seeing different shades of grey. Since, on Chalmers' interpretation of information, everything is covered in it, it follows that everything is in some degree conscious. The result? Panpsychism, a third untenable position...

 Chalmers does not actually endorse panpsychism; he just speculates about it. Do you think the idea is uninteresting? Can you not accept that if philosophers aren't allowed to speculate, they're not going to achieve very much?

Bitbucket And then, a chapter about the correct interpretation of quantum physics! What's that about, then?

 Chalmers sees a kind of harmony between his views and one of the possible interpretations of quantum theory. I have no idea whether he's on to anything, but this sort of linkage is potentially valuable, especially to philosophy, which has tended to cut itself off from contemporary science. But the point is, all these latter speculations are just that - interesting, stimulating speculations. Chalmers never pretends they're anything else. The point of the book is to get people to take qualia seriously. That's a good, well-founded project and I think even you would have to admit that the book has succeeded to a remarkable degree.

 If you ask me, Chalmers basically gives the whole thing away early on, when he says that another way of looking at the psychological/phenomenal distinction is to see them as the third-person and first-person views. Wouldn't common sense suggest that this is just a case of a single phenomenon looked at from two different points of view? It seems the obvious conclusion to me.

 But if the mind-body problem has taught us anything, it is that nothing about consciousness is obvious, and that one person's obvious truth is another person's absurdity...

Colin McGinn - cognitive closure

8 January 2004

Colin McGinn is the most prominent of the New Mysterians - people who basically offer a counsel of despair about consciousness. Look, he says, we've been at this long enough - isn't it time to confess that we're never going to solve the problem? Not that there's anything magic or insoluble about it really: it's just that our minds aren't up to it. Everything has its limitations, and not being able to understand consciousness just happens to be one of ours. Once we realise this, however, the philosophical worry basically goes away.

McGinn doesn't exactly mean that human beings are just too stupid; nor is he offering the popular but mistaken argument that the human brain cannot understand itself because containers cannot contain themselves (so that we can never absorb enough data to grasp our own workings). No: instead, he introduces the idea of cognitive closure. This means that the operations the human mind can carry out are incapable in principle of taking us to a proper appreciation of what consciousness is and how it works. It's as if, on a chess board, you were limited to diagonal moves: you could go all over the board but never link the black and white squares. That wouldn't mean that one colour was magic, or immaterial. Equally, from God's point of view, there's probably no mystery about consciousness at all - it may well be a pretty simple affair when you understand it - but we can no more take the God's-eye point of view than a dog could adopt a human understanding of physics.

Isn't all this a bit impatient? Philosophers have been chewing over problems like this quite happily for thousands of years. Suddenly, McGinn's got to have the answer right now, or he's giving up?

Anyway, it's the worst possible time to wave the white flag. The real reason these problems haven't been solved before is not because the philosophy's difficult - it's because the science hasn't been done. Brain science is difficult: you're not allowed to do many kinds of experiment on human brains (and until fairly recently the tools to do anything interesting weren't available anyway). But now things are changing rapidly, and we're learning more and more about how the brain actually works every year. McGinn might well find he's thrown the towel in just before the big breakthrough comes. A much better strategy would be to wait and see how the science develops. Once the scientists have described how the thing actually works, the philosophers can make some progress with their issues (if it matters).

There's more than just impatience behind this. McGinn points out that there are really only two ways of getting at consciousness: by directly considering one's own consciousness through introspection, or through investigating the brain as a physical object. On either side we can construct new ideas along the same kind of lines, but what we need are ideas that bridge the two realms: about the best we can do in practice is some crude correlations of time and space.

McGinn acknowledges a debt to Nagel, and you can see how these ideas might have developed out of Nagel's views about the ineffability of bat experience. According to Nagel, we can never really grasp what it's like to be a bat; some aspects of bathood are, as McGinn might put it, perceptually closed to us. Now if all our ideas stemmed directly from our perceptions (as is the case for a 'Humean' mind), this would mean that we suffered cognitive closure in respect of some ideas ('batty' ones, we could say). Of course, we're not in fact limited to ideas that stem directly from perceptions; we can infer the existence of entities we can't directly perceive. But McGinn says this doesn't help. In explaining physical events, you never need to infer non-physical entities, and in analysing phenomenal experience you never need anything except phenomenal entities. So we're stuck.


It seems to me that if there were things we couldn't perceive or infer, we wouldn't be worried about them in the first place - what difference would they make to us? If the answers on consciousness are completely beyond us, surely the questions ought to be beyond us too. Dogs can't understand Pythagoras, but that's because they can't grasp that there's anything there to understand in the first place.

Any entity which makes a difference to the world must have some observable effects, and unless the Universe turns out to be deeply inexplicable in some way, these effects must follow some lawlike pattern. Once we've observed the effects and identified the pattern, we understand the entities as far as they can be understood. If philosophers want to speculate about things that make no difference to the world, I can't stop them - but it's a waste of time.


I'm afraid it's perfectly possible that we might be capable of understanding questions to which we cannot understand the answers. Think of the chess board again (my analogy, I should say, not McGinn's). A bishop only understands diagonal moves. He can see knights moving all over the board and at every step they move from the white realm to the black realm or vice versa. He can see spatial and chronological correlations (a bit fuzzy, but at least he knows knights never move from one side of the board to the other), and both the white and black realms are quite comprehensible to him in themselves. He can see definite causal relations operating between black and white squares (though he can't predict very reliably which squares are available to any given knight). He just can't grasp how the knights move from one to the other. It looks to him as if they pop out of nowhere, or rather, as if they have some strange faculty of Free Wheel.

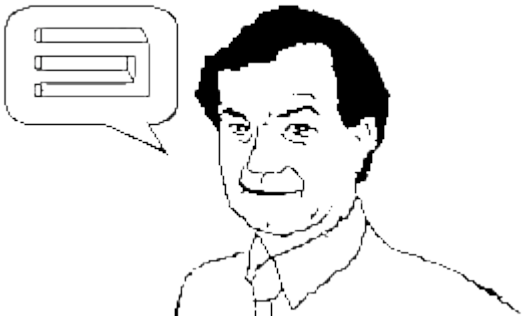

Yeah, yeah. It could be like that. But it isn't. As a matter of fact, we can infer mental states from physical data - we do it all the time, whenever we work out someone's attitude or intentions from what they're doing or the way they look. McGinn should know this better than most, given his background in psychology. Or did he and his fellow psychologists rely entirely on people's own reports of their direct phenomenal experience?

It still seems like defeatism to me, anyway. It's one thing to admit we don't understand something yet, but there is really no need to jump to the conclusion that we never will. Even if I thought McGinn were right, I think I should still prefer the stance of continuing the struggle to understand.


The point you're not grasping is that in a way, showing that the answer is unattainable is itself also an answer. There's nothing shameful about acknowledging our limitations - on the contrary. It is deplorably anthropocentric to insist that reality be constrained by what the human mind can conceive!

Roger Penrose – Non-computability

9 January 2004



Sir Roger Penrose is unique in offering something close to a proof in formal logic that minds are not merely computers. There is a kind of piquant appeal in an argument against the power of formal symbolic systems which is itself clothed largely in formal symbolic terms. Although it is this 'mathematical' argument, based on the famous proof by Gödel of the incompleteness of arithmetic, which has attracted the greatest attention, an important part of Penrose's theory is provided by positive speculations about how consciousness might really work. He thinks that consciousness may depend on a new kind of quantum physics which we don't, as yet, have a theory for, and suggests that the microtubules within brain cells might be the place where the crucial events take place. I think it must be admitted that his negative case against computationalism is much stronger than these positive theories.

Besides the direct arguments about consciousness, Penrose's two books on the subject feature excellent and highly readable passages on fractals, tiling the plane, and many other topics. At times, it must be admitted, the relevance of some of these digressions is not obvious - I'm still not convinced that the Mandelbrot Set has anything to do with consciousness, for example - but they are all fascinating and remarkably lucid pieces in their own right. 'The Emperor's New Mind' is particularly wide-ranging, and would be well worth reading even if you weren't especially interested in consciousness, while a large part of 'Shadows of the Mind' is somewhat harder going, and focuses on a particular argument which purports to establish that "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth".

 I like the books myself, mostly, but I don't find them convincing. Of course, people find a lengthy formal argument intimidating, especially from someone of Penrose's acknowledged eminence. But does anyone seriously think this kind of highly abstract reasoning can tell us anything real about how things actually work?

 You don't think maths tells us anything about the real world then? Well, let's start with the Gödelian argument, anyway. Gödel proved the incompleteness of arithmetic, that is, that there are true statements in arithmetic which can never be proved arithmetically. Actually, the proof goes much wider than that. He provides a way of generating a statement, in any formal algebraic system, which we can see is true, but which cannot be proved within the system. Penrose's point is that any mechanical, algorithmic, process is based on a formal system of some kind. So there will always be some truths that computers can't prove - but which human beings can see are true! So human thought can't be just the running of an algorithm.

 These unprovable truths are completely uninteresting ones, of course: the sort of thing Gödel produces are arid self-referential statements of no wider relevance. But in any case, the

doctrine that people can always see the truth of any such Gödel statement is a mere assertion. In the simple cases Penrose considers, of course human beings can see the truth of the statements, but there's no proof that the same goes for more complex ones. If we actually defined the formal system which brains are running on, I believe we might well find that the Gödel statement for that system really was beyond the power of brains to grasp.


I don't think that that could ever happen - it just doesn't work like that. The complexity of the system in question isn't really a factor. And in any case, brains are not 'running on' formal systems!


Oh, but they have to be! I'm not suggesting the 'program' for any given brain is simple, but I can see three ways we could in principle construct it.

One. If we list all the sensory impressions and all the instructions to act that go into or out of a brain during a lifetime, we can treat them as inputs and outputs. Now there just must be some function, some algorithm, which produces exactly those outputs for those inputs. If nothing simpler is available (I'm sure it would be) there is always the algorithm which just lists the inputs to date and says 'given these inputs, give this output'.

Two. If you don't like that approach, I reckon the way neurons work is sufficiently clear for us to construct a complete neuronal model of a brain (in principle - I'm not saying it's a practical proposition); and then that would clearly represent an implementation of a complex function for the person in question.

Three. As a last resort, we just model the whole brain in excruciating detail. It's a physical object, and obeys the normal laws of physics, so we can construct a mechanical description of how it works.

Any of these will do. The algorithms we come up with might well be huge and unwieldy, but they exist, which is all that matters. So we must be able to apply Gödel to people, too.


Nonsense! For a start, I don't believe 'inputs' and 'outputs' to human beings can be defined in those terms - reality is not digital. But the whole notion of a person's own algorithm is absurd! The point about computers is that their algorithms are defined by a programmer and kept in a recognised place, clearly distinguished from data, inputs, and hardware, so it's easy to say what they are in advance. With a brain, there is nothing you can point to in advance as the 'brain algorithm'. If you insist on interpreting the brain as running an algorithm, you just have to wait and see which bits of the brain and which bits of the rest of the person and their environment turn out to be relevant to their 'outputs' in what ways and then construct the algorithm to suit. We can never know what the total algorithm is until all the inputs and outputs have been dealt with. In short, it turns out not to be surprising that a person can't see the truth of their own Gödel statement, because they have to dead before anyone can even decide what it is!


Alright, well look at it this way. We're only talking about things that can't be proved within a particular formal system. Humans can see the truth of these statements, and even prove

them, because they go outside the formal system to do so. There's no real reason why a computer can't do the same. It may operate one algorithm to begin with, but it can learn and develop more comprehensive algorithms for itself as it goes. Why not?

 That's the whole point! Human beings can always find a new way of looking at something, but an algorithm can't. You can't have an algorithm which generates new algorithms for itself, because if it did, the new bits would by definition be part of the original algorithm.

 I think it must be clear to anyone by now that you're just playing with words. I still say that all this is simply too esoteric to have any bearing on what is essentially a practical computing problem. If I understand them correctly, both Dennett and your friend Searle agree with me (in their different ways). The algorithms in practical AI applications aren't about mathematical proof, they're about doing stuff.

 I was puzzled by Dennett's argument in 'Darwin's dangerous idea' in particular. He's quite dismissive about the whole thing, but what he seems to say is this. The narrow set of algorithms picked out by Penrose may not be able to provide an arithmetical proof, but what about all the others which Penrose has excluded from consideration? This is strange, because the ones excluded from consideration, according to Dennett, are: algorithms which don't do anything at all; algorithms which aren't interesting; algorithms which aren't about arithmetic; algorithms which don't produce proofs; and algorithms which aren't consistent! Can we reasonably expect proofs from any of these? Maybe not, says Dennett, but some of them might play a good game of chess... This seems to miss the point to me.

What I fear is that this kind of reasoning leads to what I call the Roboteer's argument (I've seen it put forward by people like Kevin Warwick and Rodney Brooks). The Roboteer says, OK, so computers will never work the way the human brain works. So what? That doesn't mean they can't be intelligent and it doesn't mean they can't be conscious. Planes don't fly the way birds do, but we don't say it isn't proper flight because of that...

 Personally, I don't see anything wrong with that argument. What about this quantum malarkey? You're not going to tell me you go along with that? There is absolutely no reason to think quantum physics has anything to do with this. It may be hard to understand, but it's just as calculable and deterministic as any other kind of physics. All there really is to this is that both consciousness and quantum physics seem a bit spooky.

 It isn't conventional, established quantum physics we're talking about. Having established that human thought goes beyond the algorithmic, Penrose needs to find a non-computable process which can account for it; but he doesn't see anything in normal physics which fits the bill. He wants the explanation to be part of physics - you ought to sympathise with that - so it has to be in a new physical theory, and new quantum physics is the best candidate. Further strength is given to the case by the ideas Stuart Hameroff and he have come up with about how it might actually work, using the microtubules which are present in the structure of nerve cells.

 They're present in most other kinds of cell, too, if I understand correctly. Microtubules have perfectly ordinary jobs to do within cells which have nothing to do with thinking. We don't understand the brain completely, but surely we know by now that neurons are the things that do the basic work.

 It isn't quite as clear as that. There has been a tendency, right since the famous McCulloch and Pitts paper of 1947, to see neurons as simple switches, but the more we know about them the less plausible that seems. Actually there is some highly complex chemistry involved. Personally, I would also say that the way neurons are organised looks very much like the sort of thing you might construct if you wanted to catch and amplify the effects of very small-scale events. One molecule - in the eye, one quantum, as Penrose points out - can make a neuron fire, and that can lead to a whole chain of other firings.

 At the end of the day, the problem is that quantum physics just doesn't help. It doesn't give us any explanatory resources we couldn't get from normal physics.

 That's too sweeping. There are actually several reasons, in my view, to think that quantum physics might be relevant to consciousness (although these are not Penrose's reasons). One is that the way two different states of affairs can apparently be held in suspense resembles the way two different courses of action can be suspended in the mind during the act of choice. A related point is the possibility that exploiting this kind of suspension could give us spectacularly fast computing, which might explain some of the remarkable properties of the brain. Another is the special role of observation - becoming conscious of things - in causing the collapse of the wavefunction. A third is that quantum physics puts some limits on how precisely we can specify the details of the world, which seems to militate against the kind of argument you were making earlier, about modelling the brain in total detail. I know all of these are open to strong objections: the real reason, as I've already said, is just that quantum physics is the most likely place to find the kind of new science which Penrose thinks is needed.

 I don't see it. It seems to me inevitable that any new physics that may come along is going to be amenable to simulation on a computer - if it wasn't, it hardly seems possible it could be clear enough to be a reasonable theory.

 In other words, your mind is closed to any possibility except computationalism. Consciousness seems to me to be such an important phenomenon that I simply cannot believe it is something just 'accidentally' conjured up by a complicated computation...

Gerald Edelman - selection and re-entry

10 January 2004

Gerald Edelman's theories are rooted in neurology. In fact, he insists that this is the only foundation for a successful theory of consciousness: the answers are not to be found in quantum physics, philosophical speculation, or computer programming.

The structure of the brain is accordingly a key factor. The neurons in the brain wire themselves up in complex and idiosyncratic patterns patterns during growth and then experience: no two people are wired the same way. The neurons do come to compose a number of structures, however. They form groups which tend to fire together, and for Edelman these groups are the basic operating unit of the brain. The other main structures are maps. An uncontroversial example here might be the way some sheets of neurons reproduce the pattern of activity on the retina at the back of the eye (with some stretching and squashing), but Edelman sees similar strucures as applying much more widely, and mapping not just sensory inputs, but each other and other kinds of neuronal activity. The whole system is bound together by re-entrant connections, sets of paths which provide parallel connections from group A to Group B and Group B back to Group A.

The principle which makes this structure work is Neuronal Group Selection, or Neural Darwinism. Some patterns are reinforced by experience, while many others are eliminated in a selective process which resembles evolution. Edelman draws an analogy with the immune system, which produces a huge variety of random antibodies: those which link successfully to a foreign substance reproduce rapidly. This explains how the body can quickly produce antibodies for substances it has never encountered before (and indeed for substances which never existed in the previous history of the planet): and in an analogous way the Theory of Neuronal Group Selection (TNGS) explains how the brain can recognise objects in the world without having a huge inherited catalogue of patterns, and without an homunculus to do the recognising for it.

The re-entrant connections between neuronal groups in different parts of the brain co-ordinate impressions from the different senses to provide a coherent, consistent, continuous experience; but re-entry is also the basic mechanism of recategorisation, the fundamental process by which the brain carves up the world into different things and recognises those it has encountered before. The word recategorisation is potentially confusing here for two reasons: first, it is not to be taken as implying the existence of a prior set of categories: in fact, every act of recognition modifies the category; nor is it meant to suggest any parallel with Kant's categories, which limit how we can understand the world. Very much the reverse, in fact.

Edelman attaches great importance to higher-order processes - concepts are maps of maps, and arise from the brain's recategorising its own activity. Concepts by themselves only constitute primary (first-order) consciousness: human consciousness also features secondary consciousness (concepts about concepts), language, and a concept of the self, all built on the foundation of first-order concepts.

The final key idea in the theory, another one with a slightly misleading name isvalue, a word used here to describe inbuilt tendencies towards particular behaviour. These forms of behaviour may be driven by what we value in a fairly straightforward sense - seeking food, for

example, but they also include such inherent actions as the hand's natural tendency to grasp. Edelman seems to think that, like a computer, if left to itself the brain might sit and do nothing. It's the value systems which supply the basic drives. This sort of set-up has been modelled in a series of robots rather cheekily named Darwin I to IV.

Edelman is emphatically opposed to the idea that the brain is a computer , however.


Being anti-computationalist but using robots to support your theory seems a little strange. It needn't be strictly contradictory, of course, but it does expose the curious fact that while Edelman insists the brain is not a computer, all the processes he describes seem perfectly capable of computerisation. He gives two reasons for not considering the brain a computer: one, that individual brains are wired up in very different ways; and two, that reality is not an orderly program feeding into the brain. Neither of these is convincing. Computers can differ enormously in physical detail while remaining essentially the same - how much similarity is there between a PC and a model Turing machine, for example - and wiring differences between brains might perhaps count as differences in pre-loaded software rather than anything more fundamental. Certainly reality does not structure itself like a program, but why should it? The analogy is with data, not with the program: you have to think of the brain as a computer which has its software loaded already and is dealing with the data coming down a wire from cameras (eyes), microphones (ears), and so on. I see no problem with that.


The argument is a bit more specific than you make out. Edelman points out that the selective processes he has in mind have an unusual feature he calls 'degeneracy' (I'm not quite sure why). Degeneracy means that the same output can be reached in a whole range of different ways. This is a feature of the immune system as well as mental processes, but it doesn't look much like an algorithm. Of course there are other arguments against considering the brain a computer, but I think Edelman's main point is that to deal with reality, you have to be able to arrange the streams of mixed-up and ever-changing data from the senses into coherent objects. Your computer with a camera attached finds this impossible except in cases where the 'reality' has been made artificially simple - a 'toy world' - and the computer has been set up in advance with lots of information about how to recognise the objects in the 'toy world'. I know you're going to tell me that great strides have been made, and that you only need another couple of decades and it'll all be sorted.


I wasn't, though it's true. I was just going to point out again that, however difficult it may be to digest reality, Edelman gives us no definite reasons to think computers couldn't do it; his robots even demonstrates some aspects of the methods he thinks most likely. But never mind.You expect me to attack Edelman just because he and Searle have spoken favourably of each other: but actually I've got nothing much against him except that I think he's misunderstood the nature of computationalism. Just because we haven't got USB ports in the back of our heads it doesn't mean brain activity isn't computable.

As for that bit about 'degeneracy', I don't see it at all. Imagine we had a job we wanted done by computer - we call in a hundred consultants to tender for the project. They'll find a hundred different ways to do it. Even if we set aside most of the possible variation - whether to use PCs, Macs, Unix boxes or what, Java, C++, visual Basic or whatever. Even if we assume the required

outputs are narrowly defined and all the tenderers have to code in bog-standard C, there'll be thousands of variations. So I reckon computers can be degenerate too...

 I don't expect you to attack Edelman at all. As a matter of fact, I'm not an unqualified admirer myself. Take his views on qualia. The temptation for a scientist is always to miss the point about qualia and end up explaining the mechanics of perception instead (a different issue) Edelman, in spite of his scientist bias, is not philosophically naive and a lot of the time he seems to understand the point perfectly. But in 'A Universe of Consciousness' he swerves at the last minute and ends up talking about how the neurons could map out a colour space - which might be interesting, but it ain't qualia. Perhaps his co-author is to blame.

However, I'm with him on the computer issue. Edelman's views about selection illustrate exactly why computers can't do what the brain does. I think his ideas on this are really important and have possibly been undersold a bit. The thing about programming a computer to deal with real situations is that you have to anticipate every possible kind of problem it might come up against - but there are an infinite number of different kinds of problem. Now this is exactly the kind of issue the immune system faced: it had to be ready to deal with any molecule whatever, no matter how novel. The solution is analogous: the immune system fills your body with a really vast number of variant antibodies; your brain is full of an astronomical number of different neuronal patterns. When the problem comes along, even a completely novel one, you're going to have the correct response waiting somewhere: and the one that matches gets reinforced and reused. Edelman called this a Darwinian process: it isn't really (hence Crick's joke about it really being 'neural Edelmanism'): the remarkable thing is, it might be better than Darwinian in this context!

 Anything's better than Darwin to you, up to and including spontaneous generation and Divine Creation.

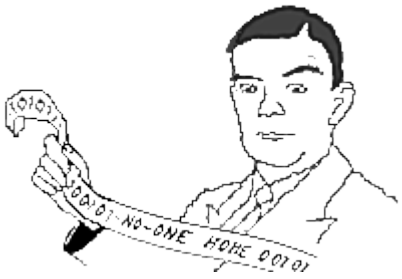
 Nonsense! But, honestly. It's not particularly original to suggest that the mind might use selective or Darwinian mechanisms, (or be infected with memes evolved in the memosphere) but normal Darwinian selection is just obviously not the answer. When we confront a sabre-toothed tiger or think what to say to a question in an interview, we don't start by copying some earlier response, try it out repeatedly and gradually refine it by random mutation. We don't even do that in our heads, normally. 99% of the time, the response is instant, and appropriate, with nothing random about it at all. It's a bit easier to understand how this could be so on the Edelman theory, because some reasonably appropriate responses could already be sloshing around in the brain and the best one could be reinforced very quickly.

 I think you're going further on that than Edelman himself would be inclined to do. In fact, I'll give you a prediction. Eventually, Edelman himself will come round to the view that there is nothing unique about all these processes, and that while the brain may not be literally a computer, its processes are computable.

 I think not. You ought to remember what the man said himself about changes of heart - the unit of selection in successful theory creation is usually a dead scientist...

Is the Turing Era Over?

2 February 2004



Can machines think? That was the question with which Alan Turing opened his famous paper of 1950, 'Computing machinery and intelligence'. The question was not exactly new, but the answer he gave opened up a new era in our thinking about minds. It had been more or less agreed up to that time that consciousness required a special and particularly difficult kind of explanation. If it didn't require spiritual intervention, or outright magic, it still needed some special power which no mere machine could possibly reproduce. Turing boldly predicted that by the end of the century we should have machines which everyone habitually treated as conscious entities, and his paper inspired a new optimism about our ability to solve the problems. But that was 1950. I'm afraid that fifty years of work since then have effectively shown that the answer is no - machines can't think

 : A little premature, I think. You have to remember that until 1950 there was very little discussion of consciousness. Textbooks on psychology never mentioned the subject. Any scientist who tried to discuss it seriously risked being taken for a loony by his colleagues. It was effectively taboo. Turing changed all that, partly by making the notion of a computer a clear and useful mathematical concept, but also through the ingenious suggestion of the Turing Test . It transformed the debate and during the second half of the century it made consciousness the hot topic of the day, the one all the most ambitious scientists wanted to crack: a subject eminent academics would take up after their knighthood or Nobel. The programme got under way, and although we have yet to achieve anything like a full human consciousness, it's already clear that there is no insurmountable barrier after all. I'd argue, in fact, that some simple forms of artificial consciousness have already been achieved.

But Turing's deadline, the year 2000, is past. We know now that his prediction, and others made since, were just wrong. Granted, some progress has been made: no-one now would claim that computers can't play chess. But they haven't done that well, even against Turing's own test, which in some ways is quite undemanding. It's not that computers failed it; they never got good enough even to put up a serious candidate. You say that consciousness used to be a taboo subject, but perhaps it was just that earlier generations of scientists knew how to shut up when they had nothing worth saying...

Of course, people got a bit over-optimistic during the last half of the last century. People always quote the story about Marvin Minsky giving one of his graduate students the job of sorting out vision over the course of the summer (I have a feeling that if that ever happened it was a joke in the first place). Of course it's embarassing that some of the wilder predictions have not come true. But you're misrepresenting Turing. The way I read him, he wasn't saying it would all be over by 2000, he was saying, look, let's put the philosophy aside until we've got a computer that can at least hold some kind of conversation.

But really I'm wasting my breath - you've just got a closed mind on the subject. Let's face it, even if I presented you with a perfectly human robot (even if I suddenly revealed that I myself had been a robot all along), you still wouldn't accept that it proved anything, would you?

 Your version of Turing sounds relatively sensible, but I just don't think his paper bears that interpretation. As for your 'perfectly human' robot, I look forward to seeing it, but no, you're right, I probably wouldn't think it proved anything much. Imitating a person, however brilliantly, and being a person are two different things. I'd need to know what was going on inside the robot, and have a convincing theory of why it added up to real consciousness.

"Picture: No theory is going to be convincing if you won't give it fair consideration. I think you must sometimes have serious doubts about the so-called problem of other minds. Do you actually feel sure that all your fellow human beings are really fully conscious entities?

 Well...

Welcome Our New Zombie Overlords

15 February 2004

A completely new idea in the field of consciousness doesn't come along all that often, but I think Professor Joel Marks of the University of New Haven may just have come up with one. In the latest issue of Philosophy Now he asks whether consciousness might not actually be a disadvantage, just as (in his eyes) having colour in the comic strips every day instead of black and white represents a backward step.

He's talking about phenomenal consciousness, of course; qualia, the real redness of red or blueness of blue. We often do things best, he points out, when we do them automatically: when we stop to become really conscious of how things look or sound, we make mistakes. Perhaps when Mary finally emerged from her black-and-white room, seeing colour for the first time would merely make her trip over the carpet.

But if that's true, then zombies (hypothetical people who function normally but have no conscious experience, just colourless registration of data) would have a consistent edge over the rest of us. They would be bound to take over the world. Just a minute, now - you don't think that they could already...

Marks is just indulging his penchant for fabulation, but there is a serious point lurking here somewhere, and once again it is clearly that qualia are a load of rubbish. The orthodox theory of qualia requires that they make no difference to the practical functioning of a human being, so that zombies without them are perfectly imaginable. But there are two reasons why that can't be so.

First, everything requires energy, however small an amount. Qualia have been likened to the humming noise that the computer makes when you switch it on, or the whistle on a steam engine. Neither, it is claimed, affect the operation of the machine. In the case of the computer, it looks as if this might be true, but if you try to make one which is otherwise identical but doesn't make the noise, you'll soon find out it isn't so. The whistle is an even worse example, because it does have a small direct effect on the operation of the engine. If you blow the whistle long enough and hard enough, the locomotive will actually begin to lose power. It follows that zombies couldn't be exactly like normal people.

Second, everything has consequences. If there could be people with qualia and people without qualia, one sort would have an advantage. Either the qualia would help you spot food and predators, or, as Marks suggests, they would be an unwelcome distraction. Either way, one sort of person would have died out long ago.

Oh, that doesn't follow at all! Think of sickle-cell anaemia and malaria. The anaemia is a bad thing, but worth having where malaria is rife and untreatable, because it gives some immunity. But that doesn't mean everyone in the relevant areas has anaemia. A balance is reached. The same might be true of zombies. Maybe they're good at say, being accountants and running large corporations, but rubbish at thinking of new ideas and telling jokes (or vice versa).

If there are too many accountants, the jokers will breed more successfully, but if everyone is laughing and can't add up, the numerate people will start to do better. A balance is struck and we end up with some zombies and some 'normal' people in the population.

Nice to hear you argue that qualia must be significant, though...

 If there were such things they would have to be: in fact, however, we are all zombies in the sense of not having real qualia, and always were zombies; or if you don't like looking at it that way, you can say we've all got qualia, but that qualia are just a normal psychological phenomenon with causes and effects just like everything else. Take your pick - basically we're better off just forgetting the whole mess.

Cognifiction

1 March 2004

Looking at Dan Lloyd's recent book "Radiant Cool" has reminded me once again how consciousness gurus and novelists have a kind of fascination with each other, or at least, with each other's work.

Lloyd himself must be very aware of this - I believe one of the courses he teaches is "philosophy in literature". The first part of his book is a kind of detective story, albeit dominated by cognitive problems: it's followed by a more straightforward account of his own theory of consciousness (which we'll look at another time). Lloyd is certainly engaged in a curiously personal way with his work: he puts himself into his novel as a character, and on his website he includes a bogus forum where Miranda Sharpe, the heroine of the story, supposedly communicates with him.

A kind of interplay between the study of cognition and fiction isn't such a new development, of course: it goes back at least to the days when the James brothers (Henry and William) set up a kind of pincer movement around the nature of conscious human experience: on one side, sensitive, reflective novels, on the other, pioneering works of psychology. The relationship between the two projects and the two brothers sounds as if it might be the basis of an interesting story itself.

It comes as no surprise, of course, to see that Dan Lloyd has written a "cautionary critique" about William James. It's interesting to observe how highly William James is rated these days by many cognitive scientists. He is, I suppose, about the only eminent figure from the history of psychology who isn't committed to some theory which has since been discredited. Most of the grand theories are in ruins these days, so the Jamesian viewpoint seems modern, almost prescient. When behaviourism was in full flood, he must have seemed just a boring Victorian.

In fact, it may be that at one time his influence on literature was greater than his impact on psychology - it was William James, after all, who first came up with the 'stream of consciousness', a theory which inspired one of the great literary movements of the twentieth century and retains an influence today. Authors such as James Joyce and Virginia Woolf put unmediated consciousness at the centre of their narratives. (Or at least, they tried. You could see the stream of consciousness in literature as comparable to earlier efforts to write 'the way people really speak' - the Lyrical Ballads of Wordsworth and Coleridge, for example - which in the end merely established new literary conventions). It comes as something of a shock to discover (as highlighted by Bernard Baars in a recent JCS article) that BF Skinner, the leading behaviorist, was himself an enthusiastic reader and author of stream of consciousness novels (he also chose the medium of fiction for one of the main expressions of his theory, 'Walden II') . This is a man, after, all, whose whole professional life was devoted to denying not just the importance, but the very existence, of internal conscious states. For him to be interested in the stream of something he didn't believe existed seems to some to imply hypocrisy, as though it had suddenly been discovered that Richard Dawkins had actually been going to confession every week.

 It's not the same, though, is it? It's more as if an eminent physicist turned out to like writing fairy stories - hardly worth a raised eyebrow really. You're a bit unfair to Skinner, too, if you ask me - it wasn't that the point of his life was to deny consciousness, he just didn't think science should waste time on waffly stuff about how people they thought they felt while there were good objective experiments waiting to be done.

Baars suggests that there is a strong connection between the novels and the cognitive theory - that failure with the novels may have helped motivate Skinner's antipathy to the whole idea of consciousness. But I grant that it's possible to be a behaviourist scientist who just happens to like novels, more or less the way you might happen to like golf or Haydn. That may be how it is in a few cases. Colin McGinn, for example, had a novel published which was not, so far as I know, a vehicle for his philosophical opinions (He has also, as he points out in his recent autobiographical book 'The making of a philosopher', appeared as a character in a novel by Edward St Aubyn).

Generally, though, the fiction produced by the consciousness experts tends to be illustrative of their theories. There is a noticeable tendency for the authors of books on consciousness to include short stories in their work. I don't mean the kind of elaborate examples quoted on these pages as 'bedtime stories' - Mary the colour scientist and her like, though these too have a curiously literary quality - but actual distinct stories. Roger Penrose brackets "Shadows of the Mind" with a short story for example. John McCarthy includes a short story 'The Robot and the Baby" on his website ('Should I publish it conventionally?' he asks). At the end of 'Brainstorms' , Dennett includes the story of what happened to him when his brain was removed and his body sent to defuse a nuclear reactor.

Dennett's view of the self as a centre of narrative gravity obviously gives fiction a special relevance. He notes, in *"Consciousness Explained",* how David Lodge's fictional character Robyn Penrose (Penrose?) reached, by an entirely different intellectual route, conclusions about the reality of the self which resemble Dennett's own. After all, he concludes "*we are both, by our own accounts, fictional characters of a sort, though of a slightly different sort*."

In *"Thinks",* David Lodge gave the consciousness scene a professional literary treatment, including parodies of a couple of the 'bedtime stories' in the style of famous authors (including, perhaps inevitably, Henry James). Apart from being a very readable novel, "Thinks" gives quite a good overall impression of the field, including a realistic sense of the yawning gap between ambition and achievement in certain areas.

Another serious contemporary novelist who has dipped his toe into consciousness theory is Ian McEwan. In 'Enduring Love' he notoriously invented a mental syndrome - 'de Clerambault's' - which nearly found its way into the academic literature as a medical reality. His hero reads an account, supposedly in a 1904 letter to Nature, of manipulative behaviour on the part of a dog. This dog is said to have got its master out of his favourite chair by the cunning ruse of scratching at the door, only to treacherously take his place. McEwan's hero points out how easily the dog's behaviour can actually be explained in the simplest behaviouristic terms, without requiring a marvellous canine ability to plan ahead and read the minds of human beings. How amusing, he says, that the appeal of a good story can lead people to accept arguments which are scientifically worthless. I haven't succeeded in tracking down the reference, but I believe this anecdote, and the related argument, don't in fact come from an old copy of Mind, but from a paper by Dennett .

The field of consciousness studies is a meeting point of many disciplines, of course. At the one end of a kind of spectrum we have engineers and programmers; somewhere in the middle are neurologists and psychologists, and at the other end the philosophers. Perhaps after all the spectrum continues with novelists? I think at the moment I'd have to say that the novelists do a better job of writing cognitive science than the cognitive scientists do with fiction.

Lucy In Disguise

15 March 2004


Steve Grand has written another book, “Growing up with Lucy” about his long-term robot-building project: partly, it seems, because he needs the money, but also because he has some interesting progress to report.

Grand, as you probably know, became famous as the brains behind the computer game ‘Creatures’, which used a far more sophisticated version of ‘artificial life’ than any of its predecessors or competitors (or successors, I suspect). In some ways, however, Grand represents an older tradition of computer games design. There was an era when computers were readily available but processing power, storage, and graphics were still severely limited. In those days it was relatively normal for an individual with a good idea to run up a world-beating game alone in his bedroom or garage. When better graphics came along and computer games started emulating feature films, all that changed. But it is in the same sort of spirit that Grand has set out to build a better robot, believing that one good man with a sound idea can achieve more than a vast instituion governed by committees. He derides those who claim that the geometric increase in computer power predicted by ‘Moore’s Law’ means that true machine consciousness is only a matter of time.

What exactly is he trying to achieve? Not a human level of consciousness, though he appears to have set his sights on an orang-utan, surely among the most thoughtful and reflective of animals. He believes that building a physical robot, and one which works in the same way as a real animal (its voice generated by real mechanical mouth/throat action rather than a chip, for example) is the most enlightening way to carry out this exercise, and he has some interesting ideas about how neural” systems might be set up in such a way that they structure and organise themselves merely through experience.


I don’t really know why someone with Grand’s programming background is insisting on building a physical robot rather than simulating the whole thing. So long as it was a sufficiently scrupulous simulation, I don’t see any issues of principle. Having to actually build the mechanical and electronic apparatus means he might easily get held up by practical engineering problems which are really nothing to do with the key issues.

There seem to me to be a few contradictions (or tensions, at least) in what he says. He says it’s no good building bits of an organism in the hope of later patching them together - you have to have the whole thing for it to work properly. Yet he’s missing out the higher levels of consciouness, which are surely deeply implicated in vision, for example (Grand himself insists that vision and other mental processes are a mixture of bottom-up and top-down pathways, which he refers to as ‘Yin’ and ‘Yang’). He scorns AI gurus who think they can get to consciousness by gradually improving what they’ve already got - he says it’s like trying to get to the Moon by gradually getting better at jumping - but at the same time he seems to think that a system of servos for a model glider gives him the beginning of a path towards genuinely mental

processes. Worst of all, though, he denies trying to produce real human-level consciousness, but puts a grotesque rubber mask, red wig, and dummy second eye on his robot, and constantly refers to it as his daughter.

 I knew you were going to take this line! Calling Lucy his daughter is really just a joke, for heaven's sake, but it does have the value of being provocative, helping to attract attention and (perhaps) funding. And I think it's reasonable to claim that it helps him if he thinks of Lucy in personal terms - it sets the right context.

 That line of reasoning reminds me vividly of a book I once read by an old-school ventriloquist. He insisted that you should always refer to the dummy as your 'partner', and treat it as a person, if you wanted your performance on stage to have real conviction. The difference is that ventriloquists admit they are trying to create an illusion! I'm particularly unhappy about that second eye. It actually seems to be a ping-pong ball or something, and looks nothing like the functioning camera eye. But on the dust-jacket of the book, it has been photoshopped to look exactly like the camera eye, reducing the repellent Frankenstein quality which I'm afraid Lucy certainly has. Retouching your pictures like this is perilously close to the line that separates window-dressing from actual deception.

The point about attracting attention is all very well, but as well as getting publicity the mask raises suspicions which actually harm Grand's credibility. Without it, people may think, it would be all too apparent that what Grand describes as one of the most advanced research robots in existence could also be described as a camera with non-functional arms. Of course, it may be that both descriptions are valid...

 I'm sorry, I'm just not going to waste any more time arguing about presentation when we could be talking about Grand's actual ideas.

He rejects the idea that there is a straightforward 'algorithm of thought' and also most current connectionist models, which he says bears no resemblance to real neural structures. He does say that he thinks you need a robot complete enough to tackle something other than 'toy' problems, but I don't accept that that conflicts with his approach to Lucy, who - alright, which - is just a test-bed for ideas anyway.

His working hypothesis is that there is a basic standard pattern of neuronal organisation which is used over and over again for slightly different functions in different parts of the brain. So instead of bolting together a set of incompatible specialised modules, the task is to find a single 'component' that can be endlessly re-customised. Basically, this comes down to neural maps which he thinks can carry out co-ordinate transforms, thereby controlling behaviour in the same sort of way as - yes - servos on an automated model glider. The point here is that the servos effectively try to make a particular quantity - flap angle, rate of roll, angle of bank, and so on up the hierarchy - match the one specified by a higher-level controller. His argument is not the naive one you suggest, but he does point out that if you had an automatic glider which had maximum altitude as its target value, there would be an analogy with a 'desire' or 'intention' on the part of the glider to stay as high as possible.

 Yes - on the same kind of level as the thermostat's 'desire' to keep the room at the right temperature, or for that matter, the 'intention' of a broken glider to plummet crazily to earth.

 This is about as close as Grand gets to the big issues of consciousness. Somewhere near or at the top of conscious organisation he speculates that there is a map which sets out a kind of ideal target state for the organism. This could be seen as representing its desires or drives; another of his angles is that a key mental task is predicting events; bringing the prediction map and the desire map into line could be what it's all about. He remarks that a saccade, an automatic jumping of the eyes towards a sudden movement or other point of interest in the visual field, is always interpreted as a conscious decision, with the strong implication that other forms of consciousness only 'seem' to be in control. Taken together with scepticism about free will, I think this gives us a fair idea of his philosophical position, even without further discussion. That's not to say there isn't a lot of interesting amd relevant stuff here, though. Grand describes the neuronal structure of the brain and points out how half or more than half the signals in sensory areas are actually going in a top-down direction; perception, he suggests, is anything but passive. He reckons there is a kind of tripod structure of paired up and down paths. He has some rather speculative ideas about how neurons might spontaneously organise themselves into 'pinwheel' structures that help to identify edge orientation and he describes a method which the brain just might be able to use to identify shapes regardless of size and orientation. All this is pretty speculative, but it has a new and promising feel.

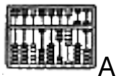
 I don't know about that. It seems to me some of this stuff is unlikely (there are a few rather serious problems about mental representations of people's desires which don't arise in the case of automatic glider-steering systems) and the rest is not very original. The idea that neuronal maps could do co-ordinate transforms, and thereby make eyes saccade towards a point of interest, and arms reach out towards it, was covered by Paul Churchland in a fairly well-known paper in Mind in 1986 - 'Some Reductive Strategies in Cognitive Neurobiology', where he also suggested the same kind of maps might have other interesting uses. But saccading eyes are a favourite trick of AI robots - Rodney Brooks' Cog, for example. I'm sorry to come back to this point again, but it's hard not to feel that eyes that follow you round the room are popular because they make the robot 'seem alive' in the same way as Lucy's mask. It looks as if the AI people are so tired of getting nowhere that they are tempted towards shallow but crowd-pleasing tricks just to get some popular approval. Look at Kismet, another robot from Brooks' laboratory, mainly the work of Cynthia Breazeal: it has big eyes, lips, and other facial features which it uses to interact with people, gurgling at them, looking them in the eye, and so on. It's supposed to have internal states analogous to emotions, and it apparently requires fifteen different computers, but frankly it seems to me to bear an embarassingly clear resemblance to a Furby.

 Scoff all you like: I think an edge of eccentricity is part of Steve Grand's appeal. You never know quite what he might come up with.

 I agree that the ‘mad inventor’ personality has a certain charm, and perhaps a certain usefulness. It has its downside too, though. I see that Grand has funding as a NESTA ‘Dreamtime’ Fellow . It’s nice to be called a dreamer in some ways, but it’s an ambivalent kind of compliment.

The Good, the Bad, and the Conscious

2 April 2004

 A team led by Uri Hasson at Tel Aviv University used an fMRI scanner to monitor the brains of a series of subjects while they were each shown the same 30 minutes of 'The Good, the Bad, and the Ugly'.

"We reasoned that such rich and complex stimulation will be much closer to ecological vision relative to the highly constrained visual stimuli used in the laboratory", he explained.

The results showed a remarkable consistency - the same film evoked just the same patterns of neural activity in all the subjects. In itself this might not seem extraordinary, until you consider the implications - this is strong evidence that the experience of watching the film is the same for everyone. I think many people would have expected that individuals' different neural wiring would influence their response and cause them to process the images in quite different ways. Well, it seems they don't. So what does that tell us about private, unique, incommunicable qualia?

 Nothing, of course. You can't read anything into these results. We already knew that the actual patterns on the retina get relayed right across the brain with only a small amount of topological distortion. I'd bet that most of what the Tel Aviv people picked up was a direct echo of what was on the screen. In the second place, watching a film is an inherently passive business. When you follow a film, you automatically stop thinking about anything else. If the subjects had been asked to watch real people enacting a situation in which they were, or could be, involved, it would have been different.

 Well, I agree there are limits to what you can read into this particular project, but in my view, it puts the writing on the wall. Qualia are going the way dreams went. For hundreds of years dreams were a deep philosophical mystery; then we discovered that everybody dreams during REM sleep, and that to some extent the content of the dreams can be influenced by outside stimuli. Now all this was pure empirical science, based on waking the subjects up and asking them what they had just been dreaming. In principle, it doesn't touch the philosophical problem at all. The philosophers should have been in there asking how the scientists knew, or could know, that reports given after waking truly reflected dreams which occurred while asleep. Couldn't it be that the act of waking people up from REM sleep caused a false memory of a preceding dream they'd never actually had, for example?

But in fact, I don't remember any significant resistance along these lines. The truth is, once the gap in our scientific knowledge had been filled, the residual philosophical point no longer seemed important. You could still be philosophically doubtful about, say, whether people dream every night, but only in the same kind of way you could be doubtful about whether anyone else has a mind at all - nobody doubts such things in practice.

Within a few years, science is going to offer direct unequivocal data showing that the red you see is the same as the red I see. That won't dispel the philosophical "hard problem" of qualia, strictly speaking - but if the philosophers want to maintain the view that it's a real issue, they'd better man the barricades now. Otherwise, people are just going to stop talking about qualia - and not before time!

Or perhaps it will be like artificial intelligence, where the first few steps were easy and then the ground suddenly dropped away from beneath the researchers' feet. Perhaps reading people's minds from brain scans is going to be a bit more difficult than you think.

 I don't say it's going to be easy. But I've got evidence on my side - all you've got is hope. Or should that be dogma?

The Final Frontier

15 April 2004


It's not uncommon for philosophers of mind to refer enviously or scathingly to the transporter they used to have on board the U.S.S. Enterprise, which could dematerialise people in one place and then reconstitute them with exquisite accuracy in some other, distant place. I assume that the way this worked was that the transporter produced a comprehensive physical description of say, Kirk, which was then sent off to the destination and somehow used as the basis of a perfect reconstruction. I'm not sure modern physics allows this sort of thing, though I've seen articles which suggest it does, at least on a small scale.

Whatever the physics, the device raises many practical questions. Why didn't they just transport to places, instead of using a spaceship? Or if that was impossible, why not store crew and equipment as data, and reconstitute them only when needed? Why did they use photon torpedoes when a transporter could presumably have scrambled the atoms of any Klingon ship at the press of a button? I'm sure there are good answers to all these questions, but the one that attracts the attention of philosophers is: was the person who arrived on the planet's surface the same James T. Kirk who stepped into the transporter aboard the Enterprise?


Of course it would have been the same person! Think of this - every molecule in Kirk's body gets changed time and again through the normal processes of human metabolism. When he gets reconstructed on the planet's surface, it's really no different- he's just put together from a different set of elementary particles. If the actual particles were what mattered, his identity would be changing all the time as his cells rebuilt themselves from the new molecules derived from his food. When we're talking about identity, it's the form and causal relations that matter, not the actual physical stuff - and the transporter reproduces them perfectly.


That's all very well, but consider this: if the transporter malfunctions, it can produce two copies of Kirk. Now if there's one thing that's certain about personal identity, it's that each person only gets one - so how can there be two? For that matter, Kirk's descendants could pick up the echoes of the transporter beam from deep space long after he's dead and reconstruct a dozen copies - they can't all be him, can they?

> PH: Excuse me interrupting, chaps, but can I introduce another idea? You both seem to be atheists, but most people have always attributed personal identity to an immaterial soul or spirit. Could it be that the question is essentially whether the transporter moves Kirk's soul along with his body?


I am not an atheist. In my view, faith and science are different levels of discourse, rather than contradictory accounts - but they do ultimately address the same world. We know that physical events can have spiritual consequences - to take a crude example, a blow on the head

can effectively separate your soul from your body. I therefore don't have any problem with the idea that if Kirk's physical body were moved by the transporter, his soul would follow. What I challenge is the view that what the transporter does is transport, rather than duplicate.

 So once again the spiritual hypothesis turns out to be null - adding nothing to the physical account. But look, in your example, why shouldn't all the copies be Kirk? I'm not suggesting they go on being the same person after the split, but they all have an equal claim to Kirk's causal history - they've all got his memories and his characteristics. So they're all Kirk: it's just that Kirk split into several people. I see that that might cause legal difficulties over his property, but so what?

The point is, in the final analysis Kirk's identity is a set of data like anything else, and data doesn't care about the details of the medium it's recorded in, so long as there is functional equivalence! On your view, Kirk's identity depends on the identity of the atoms which make up his body - but do atoms have identities? aren't they just maths, in the final analysis?

 The irony here is that you claim to be more of a materialist than me, yet you want Kirk to be some mathematical abstraction, or some kind of program, whereas I insist that he is a physical object - and that if you disperse that object, he dies. Of course he grows and changes in the normal course of events, but brute physical continuity is what matters.

If your view were true, then Kirk's consciousness would have to split in order to go with each of his duplicates - but I ask you, isn't it obvious that consciousness is essentially unitary? My consciousness can come and go, but the idea that it could split is simply unimaginable, isn't it?

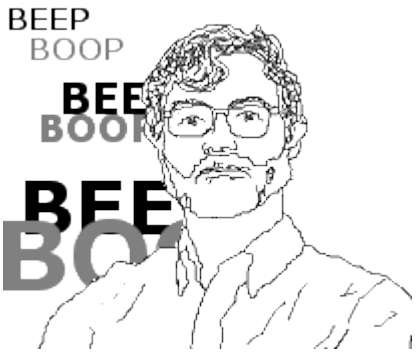
 Not to me. Anyway, some people find death unimaginable - doesn't mean it won't happen...

 But if Kirk's identity and consciousness are merely data, then they don't really have a substantial existence. If he's just data, then he exists just as much in the unreconstructed transporter signal as in the reconstruction - the fact that the machine has run up a physical instance of the data isn't that important. And actually, even the signal doesn't matter that much, since the numbers don't actually depend on the signal for their existence - they exist in some Platonic sphere, anyway. But that's true of all the possible sets of data. So on your view, there's no real difference between people who do exist and those who don't but might have...

 Frankly I think you're a little confused.

A Novel Theory

1 May 2004



I mentioned Dan Lloyd's book "Radiant Cool" earlier. As I suggested then, linkages between philosophy and literature are not exactly unknown, but Lloyd's is certainly the only book I know which splits neatly between a story in the front half and some really heavyweight stuff at the back.

The story is about Miranda Sharpe, who finds her Professor, Max Grue, slumped over his desk (unconscious?) when she creeps in to reclaim a folder. She leaves him there, only to find he has disappeared. Miranda has a series of encounters, with: Grue's class; the arch-computationalist Clare Lucid; nerdy neural networker Gordon Fescue; Porfiry Marlov the former Soviet policeman and multidimensional scaling enthusiast; the menacing Zamm and Addit, who zap sections of her brain with Professor Cronkenstein's transducer/activator; and finally, with 'Dan Lloyd'. The mystery of what happened to Grue, the origins of the Chaos Bug which is currently infecting computers everywhere, and a few thoughts about consciousness, are cleared up along the way. It's a readable narrative, with limited literary ambitions: you wouldn't have been utterly surprised if Professor Plum had turned up in the library with a length of pipe at some point.

Some of Lloyd's ideas get an exposition in 'the thrill of phenomenology', the first part of the book; a more systematic treatment, covering some additional ground appears in the second (and inevitably less readable) part, 'the real firefly'.

So what are these ideas? According to Lloyd, we have been too inclined to view people as 'detectorheads'. Seeing the main function of neurons as detection has worked very well for a number of cognitive functions, but it won't do for consciousness. Why not?

Lloyd presents a version of that old classic, the brain in a vat. If our brains were taken out of our bodies, they could still be fed with cunningly contrived signals which would give us the same experiences, phenomenologically, as we get from real life. We wouldn't know the difference. We could be having the same experience of seeing a firefly whether in fact there is a real firefly out there or not. It follows that what we actually experience consciously isn't the real world 'out there' at all (and consequently consciousness isn't a matter of detection). I think this reasoning is mistaken. It's true we can misinterpret our experiences, but that doesn't mean that if our interpretation is wrong, our experiences are not really experiences of anything external.

Consider the mad scientists feeding our brains the signals which convince us we are still moving normally in the real world. Where do they get these extremely complex signals? By far the easiest way to get suitable signals would be to read them direct from reality with a camera and other suitable equipment. But in such a case, we obviously still are experiencing the real world – it's just coming to us via the scientists' recording apparatus. The scientists could use a small model world and tiny cameras instead, but that wouldn't make a fundamental difference. We might be lulled into believing we were seeing the full-sized world, but the model is still a real, external thing. Now they could go further and use a computer model which existed only as data, they might even be able to devise a program which could generate appropriate signals without an explicit model: but however far they go along the path of abstraction we'll still be detecting

something outside our own brains. Our detectors may have been bamboozled and we may be quite wrong about what they're detecting. But they're not detecting nothing.

This doesn't invalidate the rest of Lloyd's account, however. He mentions two particularly important characteristics of consciousness – superposition and temporality – and offers a way of explaining them.

Superposition is the curious quality perceived objects have of being many different things at once. When you see a cup of coffee, you also see a container of liquid, a drink, somebody's property, a cause of insomnia, and so on – and you see it as all these things simultaneously and immediately. How can this be? This is where the multidimensional scaling come in. If we like, we can treat each characteristic of a thing as a dimension. If we have the characteristic of redness, we imagine a line stretching from not red at all to utterly red: every object can be placed somewhere on this one-dimensional line according to its redness. Then we can construct an imaginary space out of these dimensions. Each position in this space will define a different combination of qualities, and hence, a different possible object. Now there are going to be an awful lot of dimensions involved in this imaginary space – Lloyd allows for billions (I think in fact that an infinite number of dimensions is required – and to complicate matters further, some characteristics are obviously related to others, rather than being capable of arbitrary independent variation). But that's OK. The technique of multidimensional scaling apparently allows this very complex space to be boiled down to a much more comprehensible three dimensions, while preserving the distance relationships between salient points (with a certain amount of compromise). What this leaves (we hope) is a grouping of objects according to their resemblances and relationships. Hovering in the simplified 3-d space , dogs cats and fish will be close together because they are all animals; but fish will be a bit further off because they aren't mammals, and they will also be part of another group, with apples and pancakes, because they are normal food items. Now, when you recognise something, you trigger some neural equivalent of this space which means you automatically see the thing you recognised in several different potential contexts at once. As you can tell, I have some reservations about the complex space apparatus here, but it does seem there is some gleam of light in the underlying idea.

When it comes to temporality, the influence of Husserl bulks large. According to Lloyd, every experience contains within it a strong sense of both past and future – retention and protention. To those with a good Husserlian background, this may seem evident, but things just don't seem like that to me: some experiences have a strong temporal element, others don't. According to Lloyd, the neural patterns which encode each experience of a given moment also encode, in a weakened form, the moment before and the moment after. Since the moment before itself contained a reflection of the moment before that, we have a kind of nesting effect.¶To prove his point, Lloyd trained a neural network to predict the arrival of a 'boop' a fixed period after a 'beep', and then investigated it in considerable detail. It proved possible (using another neural network) to reconstruct earlier and later states of the network from any given point in the beep-boop process, a property not evident when it was simply running at random. Lloyd claims there is definite empirical evidence from scanning data that similar properties apply to the 'neural networks' in the brain. This tends to validate his view of the nature of consciousness as something suffused with temporality. Of course, using a task which is inherently about time – ie waiting the right length of time for the 'boop' – might be held to have biased the results.

Lloyd acknowledges scope for some reservations, and does not expect to have provoked the much-sought 'aha!' reaction of the reader who suddenly understands the whole thing at last.

But he feels he's made some pretty good progress. I'm not convinced he's altogether on the right track, but the book is genuinely a useful prod towards some novel ways of thinking.

A lot of related material including some 3-d models and (apparently) an abortive exchange with Miranda Sharpe, can be found on Lloyd's own website.

Bats

15 May 2004


Nagel's classic "What is it like to be a bat?" must be one of the most influential papers on consciousness of the last century, and it's still very relevant.

Nagel's aim is to launch a kind of counter-attack against physicalist arguments, which would reduce the mental to the merely physical, and which were evidently getting into the ascendant in 1974 when the paper was published. Tempting as it may be to fall back on the familiar kind of reductionist approach which has worked so well in other areas, Nagel argues, phenomenal, subjective experience is a special case. Reductive arguments always seek to give an explanation in objective terms, but the essential point about conscious experiences is that they are subjective. The whole idea of an objective account therefore makes no sense - no more sense than asking what my inward experiences are really like, as opposed to how they seem to me. How they seem to me is all there is to them. Any neutral, objective, third-person explanation has to leave out the essence of the experience. The point about conscious experience is that there is something it is like to see x, or hear y, or feel z.


Ah, 'there is something it is like' - the phrase that launched a thousand papers. Surely you realise that this is just an over-literal interpretation of the conventional phrase 'what is it like?'. To assume that the 'it' in that question represents a real thing rather than a grammatical quirk is just silly.

"Blandula" Yes, I understand your point, but Nagel's whole point is that 'what it's like' is strictly inexpressible in objective terms. So it isn't surprising that he has to resort to a back-handed way of getting you to see what he's talking about. If he could describe it straightforwardly, he'd be contradicting his own theory.

Anyway. Nagel uses the example of a bat to dramatise his case - how can we know what it is like to be a bat, from the inside?


There's a large rhetorical element in the choice of a bat. Bats have the traditional reputation of being a bit weird, and it's known that some of them have a sense we don't - echolocation. All this helps to persuade people that we can't imagine what things are like from another point of view. But if Nagel is right, it should be equally hard to see things from the point of view of an identical twin. So let's get the bats out of this particular belfry, OK?

"Blandula" Nagel's entitled to use any example he likes. He explains that he chose bats because they're close enough to human beings to leave most people in no doubt that they have conscious experiences of some kind, while far enough from us to dramatise his case. But whether you like it or not, it raises some fundamental issues. If Nagel is right, there are certain experiences - bat experiences, for example - that humans can never have. It follows that there are true facts about these experiences which humans can never grasp (although they can grasp

that there must be facts of this kind. This general conclusion about the limits of human understanding must have been part of the inspiration for Colin McGinn's wider theory that even human consciousness is ultimately beyond our understanding.

 Yes, of course, since human beings are by definition not bats, they can't have the experience of being a bat. But it does not follow that there are facts about bat experiences they can't understand. You see, actually we can know what it's like to be a bat. We can know what sizes of objects echolocation detects, and how the bat directs its ears and the stream of sound, and thousands of facts of that kind. We can know all about the kinds of information a bat's senses supply, and with the right equipment we can experience echolocation ourselves at least by proxy.

I think the worst part of the paper is where Nagel says that even if we imagine ourselves turning into a bat, that won't be any good. We're just imagining what it would be like for us to be a bat, whereas we need to imagine what it's like for a bat. This just reduces the whole thing to the trivial point that we can't stop being us. Because if we did - it wouldn't be us any more!

"Blandula" You just need to make the imaginative effort to see what he's on about. Actually, the claim being made is quite modest in some respects. Nagel himself says that his argument doesn't disprove physicalism. It would be nearer the truth to say that physicalism, the view that mental entities are physical entities, is a hypothesis we can't even understand properly

Colourless Green Gavagai

20 May 2004

Strange, really, that the best known sentence Noam Chomsky ever wrote is probably the one which wasn't supposed to mean anything. In 'Syntactic structures' (1957) he pointed out that while neither

- *Colorless green ideas sleep furiously*, or
- *Furiously sleep ideas green colorless*

means anything, we can easily see that the first is a valid sentence, while the second is not. Since neither sentence had ever appeared in any text until then, statistical analysis of language won't help us tell which of them is more likely to occur in normal discourse, he said.

Strictly, the statistical point appears to be wrong – we can, in fact, assess the relative probability of sentences which have never occurred. Be that as it may, the thing that really caught people's imagination was the grammatical but meaningless sentence. Was it really meaningless? Some thought they could see a kind of poetic meaning in it. Some people at Stanford had a competition which produced a number of poetic examples, and there is at least one other piece of verse. Resorting to poetry makes thing too easy, though. A more challenging exercise might be to reinterpret the sentence as part of a crossword clue...

> *Clue: Wow, awful colorless, green ideas sleep furiously (3,4,8)*
>
> *Solution: Gee, dire paleness.*
>
> *('Furiously' indicates an anagram of 'green ideas sleep', and the result means (more or less) the same as 'Wow, awful colorless'.)*

All right, maybe not the best crossword clue ever. Without going to those lengths, we can easily imagine that the sentence might be part of a political essay...

> *Even before the fall of the Berlin Wall, it was known that 'colorless Reds' – covert Soviet agents – had taken up places as 'sleepers' in the Dubitanian government. With the benefit of hindsight, we can see that these agents were in fact among the government's most reliable and least politicised employees. Disruption, direct action, and sabotage were far more likely to be the work of the extremist ecological factions who also infiltrated certain government departments. Once securely lodged inside the state, it seems, Communist ideology waits calmly for its chance. Colorless green ideas sleep furiously.*

'Green' has so many relatively normal uses that it lends itself to different interpretations. The sentence could be about refurbishment of a golf course, abstract painting, or something new and half-baked. But we're not limited to ringing the changes on 'green'. With the right context, we can make other words - 'idea' for example - mean something slightly different, too.

> *The Studge advertising agency was desperate to win the Kumfypillo account. The creative department decided on a 'rainbow' workshop. In these sessions, each person was assigned a color and a corresponding role. When the box of equipment was opened, however, the green badge was missing, so Jenkins, the junior member of the*

team, had to be green without a color: he also got the hardest job, which was to come up with new creative angles, a process which at Studge was called 'ideaing'. Imagine the scene; in the new chairs by the window blue and red are, respectively, critiquing and relating the concept of repose; at the front of the room yellow catalogues the properties of head-support, while standing awkwardly at the back, poor colorless green ideas sleep furiously.

Without wanting to labour the point, one can imagine an interpretation in which none of the words have their usual meanings, and all are used in a different grammatical role from the one you would expect. It's actually a bit of a challenge to hold that many novel meanings in your mind at once, but...

'This is the strangest editorial office I've ever worked in,' said John, 'I can't understand half of what people say'

'You don't have to be mad to work here, son,' observed Smith, 'we can teach you all that. Let me run you through some of the slang. Now 'color' is pictures, so 'colorless' means text, or words, as in 'Mark my colorless, son'. "Green" is for green light, means "OK", "can do" as in "I green lunch today". OK so far? Now at one time the boss had a habit of picking up some piece and saying 'But what does it mean? What are the ideas?'. So if you want to ask what somebody means, you can say 'what ideas there, then?' Now "sleep" means, "relax", "it's correct". So if someone asks me if I'm going out today, I can say "hey, sleep", meaning "certainly". One more. When we're in a desperate hurry to finish, we generally just furiously chuck in whatever stuff we've got. So "furiously" means "whatever you've got", or "anything", like, if somebody asks me what I'm drinking and I don't care, I'll just say "Oh, furiously.".

'Good grief', exclaimed John, 'Words can mean really anything.'

'That's what I'm saying,' answered Smith 'Colorless green ideas sleep furiously'.

 Gibberish. Alright, I admire your imagination, or perhaps it would be nearer the mark to say your dull persistence in wringing out permutations. But so what? Inventing a code in which each word stands for a completely different one is a futile pursuit, and tells us nothing about Chomsky.

 I'm not trying to make a point about Chomsky. What I'm actually doing is suggesting that Quine was right with his story about 'gavagai ' - the word which seemed to mean 'rabbit', but could have meant virtually anything. For any word or set of words, there really are an infinite number of possible interpretations, and it follows that the meaning is always impossible to decode with any certainty.

 Then how does anyone ever understand anyone else? Quine's view is one of those theories that even the author doesn't believe when it comes to real life.

Ah, but you see, the thing is, we don’t decode words mechanically, the way a computer would have to do it. We just see what somebody means – it’s a process of recognition, not calculation. Perhaps that has some implications for Chomsky after all.

Electromagnetic

30 May 2004

We have become used to hearing that novel theories of quantum mechanics might somehow account for consciousness. A theory which invokes only common or garden electromagnetism seems refreshingly simple by comparison - almost naively simple at first sight. But such is the hypothesis put forward by Susan Pockett in her book 'The Nature of Consciousness'. The hypothesis can be briefly stated - *'consciousness is identical with certain spatiotemporal patterns in the electromagnetic field'* .

The original inspiration for the theory apparently lies in the sense of cosmic oneness described in Hinduism and (rather an unexpected choice) Plato. Pointing out that the ancients had some handy ideas about atoms, Pockett suggests we could similarly look for ancestral wisdom as a starting point for an enquiry into consciousness. One possible clue, according to her, is that one can find in Hinduism (and Plato) the idea of a fundamental underlying unity, a universal consciousness. I don't think anyone would have argued very much if she had claimed it was a common feature of mystical experience in virtually all religions, actually. This mystical element is clearly important to her, since it is brought back in to round off the book's conclusion.

 Oh great. Good to know we're dealing with hard science again.

Well, actually we are. If you want scrupulous quotation of authoritative empirical evidence, you won't be disappointed here. The theory may be mystically inspired, but being a practical New Zealander, brought up, as she says, to believe any problem can be solved with baler twine, Pockett quickly gives it a more concrete form. This universal consciousness, she asks - doesn't that sound a bit like some kind of *field*? If consciousness were some kind of electromagnetic effect, it could be part of the universal electromagnetic field, and hence genuinely part of a cosmic unity. Now, you wouldn't want it to lose its separate existence altogether, or even leak across to other people in the form of telepathy (at least, not obviously), but that's OK. If we were talking about very low frequency fluctuations, in the 0-100 Hz range, they wouldn't propagate very far and would remain pretty much isolated, barring the kind of jolt to the brain which would have significant physical effects in any case.

Is there any reason to think that the brain might work at this kind of frequency, and that if it does, that the relevant patterns vary in a way that matches the variation of conscious experience?

Before we can consider the match with conscious states, we need to be a bit clearer about the kind of consciousness we mean. Pockett is only talking about the kind of subjective, phenomenal consciousness we almost certainly share with animals - qualia, in fact, not articulate, decision-making, reflective consciousness. She identifies three different states of consciousness - waking, dreaming, and dreamless sleep - and suggests there might be grounds for accepting another - a half-way state between waking and sleep which she identifies with the kind of meditative trance mystics go in for. It appears that EEG evidence does indeed show various wave patterns at suitable frequencies for all three (or four) of these conscious states. So far so good.

The next step is to examine the evidence for covariance between these EEG patterns and conscious sensations. It's characteristic of Pockett's agreeably down-to-earth style that smell, for once, is tackled first and gets a fair share of the attention, along with hearing and vision. These chapters form an interesting survey in their own right, though the details don't matter much from our point of view. Covariance turns out to be relatively easy to establish in simple organisms, but in human beings rather a lot of processing of the relevant data is required. Nevertheless, I think most people would be happy to accept the broad conclusion that there is, indeed, evidence for covariance in all three senses.

Covariance, of course, is not enough in itself to establish the truth of Pockett's hypothesis. Most people, as she recognises, would be inclined to say that the relationship between conscious sensations and electromagnetic patterns is due to the fact that they both arise from the same patterns of neural activity. I don't think Pockett has a knock-down argument against this one: essentially she thinks it is much easier to conceive of consciousness as 'just being' a shimmering electromagnetic field than 'just being' patterns of neural activity - she even seems to suggest that the latter view tends towards dualism - but I'm not convinced.

So why should we believe in the hypothesis? I think there are three main points. One is the 'mystic unity' argument. To find this appealing, you have to believe in some kind of cosmic unity of consciousness to begin with, of course. But even if you do, I'm not sure the theory backs it up as strongly as Pockett suggests. At one point she sums it up quite accurately by saying that the 'spots' of consciousness within the overall cosmic field are like the red spots on a spotted handkerchief, which, in a sense, confer the quality of redness on the handkerchief. But that doesn't, except in a stretched or metaphorical sense, make it a red handkerchief or unite the spots in a handkerchief-wide redness. Pockett herself has to argue for isolation of individual consciousness in order to defend herself against suggestions that if her theory were true, we should all be telepathic, or disrupted by electromagnetic events around us. I don't, ultimately, find myself tempted by Pockett's suggestion that her ideas, similarly, mean that the Cosmos itself is conscious.

The second point is a claim that the theory solves the binding problem - how conscious experience appears unified although the data from different sense organs makes its way into the brain at different times. If everything feeds into a single overall electromagnetic pattern, unity is guaranteed. This is a tempting idea, and a real prize if it worked. I think in the final analysis it underrates what we already know about the importance of neural events to sensory experiences. Since the theory only deals with qualia-style consciousness, it also leaves us with a problem on our hands about how sensory data feed into the other, cogitative form of consciousness. Nevertheless, I don't think the idea that electromagnetic effects are relevant here can be entirely dismissed.

The third, and most startling point is that if Pockett is right, it is possible in principle to recreate the patterns which constitute conscious experience without a brain at all. Conscious computers are the least of it - if this is right you can generate a conscious experience of, say, the colour orange, in empty space. Pockett presumes that if someone's brain were moved to coincide with such a floating experience, the owner of the brain would indeed 'have' that experience. If Pockett could indeed find a practical way of 'beaming' chosen experiences into someone's mind (without, presumably, the need to know any details about the particular brain involved) it would be a most dramatic vindication.

I'm not holding my breath, though - it just seems unlikely that the electromagnetic aspect of the brain could ever be so thoroughly divorced from the neural activity. This touches on a basic problem with the theory which comes out in a number of different ways. One of the objections discussed and dismissed in the book is that electromagnetic fields can't do computation. Now, as a matter of fact I think the objection, as stated, is on dubious grounds anyway. It isn't clear to me that the kind of consciousness under discussion - qualia, subjective experience - is meant, by those who espouse it, to be computational anyway. But there is, I think, a problem about causal relations in the electromagnetic theory. It seems a little odd to think of electromagnetic patterns causing other electromagnetic patterns without physical objects - neurons, in fact - playing a role somewhere. The implication is either that some compromise with the neural perspective is needed (I think there are a number of reasonable options along these lines), or that the kind of consciousness under discussion is epiphenomenal - has no causal relevance. This latter view is, of course, virtually the orthodox one among qualists, but it involves considerable difficulty.

Pockett herself published a paper in the JCS recently declaring for epiphenomenalism, though whether she means the (indefensible in my view) hard philosophical version, or the unproblematic psychological version is still, I think, open to discussion.

I'm not convinced, but pending the arrival of an electromagnetic conscious-experience-synthesising machine, I think the hypothesis at least remains among the small and praiseworthy company of ideas about consciousness which are rational, clear, and in principle testable.


Ah - excuse me, but doesn't it miss the whole point of mystical religious experience (in a typically flat-footed Western materialist way) to *explain* it as being a lot of radio waves? Isn't transcendence of the physical something to do with it...?

Ambiguous Turing

10 June 2004

It's just 50 years since Alan Turing's tragic death. The anniversary was marked in Manchester and elsewhere, but little seems to have appeared on the Internet – perhaps surprisingly, given his importance in the development of the computer.

Turing has a number of tremendous achievements to his credit. His war-time codebreaking may be the most famous; but perhaps the most important was the idea of the Turing machine, the theoretical apparatus which defined computation and computers. It had two distinct consequences: on the one hand, it dealt with the Entscheidungsproblem, one of the key issues of 20th century mathematics; on the other, it gave rise, via Turing's famous (1950) paper, to the period of intense optimism about artificial intelligence which I referred to earlier as the 'Turing era'. The curious thing is that these two consequences of the Turing machine point in opposite,

How so? The Entscheidungsproblem, posed by Hilbert, asks whether there is any mechanical procedure for determining whether a mathematical problem is solvable. The universal Turing machine embodies and clarifies the idea of 'mechanical' calculation. It is a simple apparatus which prints or erases characters on a paper tape according to the rules it has been given. In spite of this extreme simplicity it can in principle carry out any mechanical computation. In theory, in fact, it can run an appropriate version of any computer program, including the ones being used to display this page. In many respects it appears to be an entirely realistic machine which could easily be put together, but it has certain other qualities which make it an impossible abstraction. For one thing, it has to have an infinite paper tape: for another, it has to be immune to malfunction, no matter how long it runs; and most fundamental of all, it has to operate with discrete states – it must switch from one physical configuration to another without any intervening half-way stages. These characteristics mean that it is really more like a complex function than a real machine. Nevertheless, all real-world computers owe their computerhood

The clear conception of computation which the Turing machine provided allowed Turing to show that the Entscheidungsproblem had to be answered in the negative – there is no general procedure which can deal with all mathematical problems, even in principle. In fact, Turing was slightly too late to claim full credit for this result, which had already been established by Alonzo Church using a different approach,

The thing is, this result goes naturally with Gödel's proof of the incompleteness of arithmetic in the sense that both establish limitations of formal algorithmic calculation. Both, therefore, suggest that the kind of computation performed by machines can never fully equal the thought processes of human beings (however those may work), which do not seem to suffer the same limitations. Gödel seems to have interpreted his own work this way. In fact there is some reason to think that Turing initially took a similar view. Andrew Hodges has pointed out that after completing his work on the Entscheidungsproblem, Turing attempted to produce a formal logic based on ordinals. It seems to have been the idea that this new, ordinal-based work would provide the basis for the kind of 'intuitive' reasoning which Turing machines couldn't deliver – the kind human beings used to see the truth of Gödel statements. Only when these efforts

failed, it seems, did Turing look for reasons to think that machine-style computation might be good enough to deliver a real mind after all.

Looked at again in this light, the 1950 paper seems more evasive and equivocal. It is a curious paper in many ways, with its playful tone and respectful mentions of ESP and Ada, Countess of Lovelace, but it also skirts the issue. Can machines think? Well, it says, let's consider instead whether they can pass the Turing test . If they can, well, perhaps the original question is too meaningless to worry about.

But it surely isn't meaningless: it's partly because we believe that people really can think that our attitude to death is so different from our attitude to switching off the computer, for example.

It seems possible, anyway, that Turing's desire to believe that a mechanical mind was possible led him to seek ways around the negative implications of his own work. The logic of ordinals was one possibility: when that failed, the Turing Test was basically another, justifiying further work with Turing-machine style computers.

Had he lived, of course, he might eventually have changed his mind about his own Test, or found better ways of dealing with 'intuition'. We'll never know quite how much we lost when, punished for his homosexuality with oestrogen injections and expelled from further participation in Government work, he killed himself with a poisoned apple.

But it is a poignant thought that in the natural course of things he could still have been alive today.

Honderich Exclusus

21 June 2004


There are many eyebrow raising sentences in Ted Honderich's paper "Consciousness as Existence, Devout Physicalism, Spiritualism", accessible on his own web-site. Not the least surprising is the confession, half-way through, that the paper was rejected by the Journal of Consciousness Studies. The lack of reserve here will be familiar to readers of Honderich's remarkably frank autobiography "Philosopher: a kind of life".

But what's going on? Honderich, now retired, had a long and eminent career in philosophy. He is the former Grote Professor at University College, London; author of many magisterial works, notably on punishment, free will, and the justifications for terrorism. He has himself acted as editor for more philosophy books than you could conveniently shake a stick at. You really wouldn't expect him to be getting many flat rejections at this stage.


I think I can shed some light on that. I believe the JCS must have rejected the paper because it just makes no sense. I've read it carefully, and it's not that I disagree with Honderich – I just cannot make out what he is getting at. Look, he says 'For you to be conscious of the room is, it seems, for the room somehow to exist.' I'd like this sentence a lot better without the 'it seems' and the 'somehow', but those are minor quibbles. Can he really mean that the room's existence is the same thing as my being conscious of it? If so, it follows that I must be conscious of everything that exists. Which is surely nonsense. Equally, if my being conscious of the room is merely a fact about the room (that it exists), the state of my brain at the time is irrelevant. So I could have exactly the same brain state while conscious of the room as I have while I'm not conscious of it. Which is also surely nonsense.

So what can he mean? He says that what is normally, or by some people, taken to be the contents of consciousness are in fact, more or less, consciousness itself. What could that mean? I can imagine someone declaring that the contents of a book were the book (rather than any actual physical copy of the book), but how would that apply to consciousness? It seems that if you interpret it one way it becomes vacuous (the fact that consciousness has contents is what distinguishes it from unconsciousness); if you interpret it another it becomes absurd (there is no distinction between the conscious thing and the thing it is conscious of).

Honderich doesn't give us all that much help in the course of the paper. He compares his theory with hard-line materialism and with dualism (the 'devout physicalism' and 'spiritualism' of the title), and he rates it against four criteria which he seems to take as obvious, but which in fact seem rather arbitrarily chosen. None of this helps much in the basic task of grasping his meaning. At one early stage I wondered if we were heading towards some kind of idealism, but Honderich, pointing out that he is 'not mad as a hatter' says his views are nothing to do with Bishop Berkeley, and no kind of epiphenomenalism, either.


I think you have to remember that Honderich has been struggling with the mind-body problem since long before it became so fashionable. I think part of his reason for stressing existence is simply to short-circuit the argument from error which was still strong thirty years ago (actually it still crops up). According to that argument, the fact that we are sometimes wrong about our perceptions shows it's really only sense-data, or images we perceive – by stressing that true consciousness involves the existence of the perceived, Honderich rules that line of thinking firmly out of court.

I don't think the theory is quite as confusing as you maintain, but I do have a bit of difficulty deciding whether it is meant to be relativistic or absolute. Some of the things said imply that each conscious entity exists in its own perceived world, where indeed existence and consciousness coincide, but it's also a key point for Honderich that his argument makes consciousness a straightforward physical phenomenon, amenable to physical investigation. I'm not sure how these two claims can be reconciled.

We mustn't forget, of course, the possibility that Honderich has got it absolutely right, and cracked the ultimate mystery of consciousness - but that we're still too stupid to understand the answer, even when it's explained to us.

Honderich himself doesn't seem to regard the theory as the final truth, though. He claims that it has the desirable quality of explaining its own limitations – if consciousness is like this, no wonder it seems permanently mysterious – and suggests we might see merit in several different theories – pursue several in tandem. That doesn't seem to me an unappealing perspective.'

Making Up A Mind

1 July 2004

In “How Brains make up their Minds”, Walter J Freeman set out to tackle the ancient issue of free will, but he also addressed many of the other fundamental issues about consciousness and thought. The book has an unusually even balance of neurology and philosophy, with similar ideas coming into play in both fields.

Freeman uses some familiar terms from philosophy in rather unusual ways. For him, intentionality does not mean “aboutness” in the way it generally does to contemporary philosophers. Instead, it means the property of being directed towards some object or goal. So in his eyes, the food-seeking behaviour of simple organisms displays intentionality even though there is no question of their having plans or acting deliberately. In his view, this is Thomas Aquinas’s original meaning, and a key foundation for consciousness. Aquinas is credited with a number of important insights which Freeman has incorporated into his own views.

‘Meaning’ also has a special sense in Freeman’s account, quite distinct from simple information. Freeman speaks of meaning in what sound at first like worryingly poetic or metaphorical terms, but the point is really a matter of context. Meaning, in Freeman’s sense, is given to mere information when it is set in the context of an individual mind, with all its multiple life experiences, history and characteristic quirks. This matches his views about the neuronal operation of the brain, where rather than discrete bits of data working their way through a program, he sees a mathematically chaotic pattern of activity in which the whole system comes to bear. Each brain has its own individual pattern of basic activity which provides a unique context in which meanings develop. It follows that meanings are, strictly, unique to particular individuals, and in stark contrast to Putnam’s famous doctrine, meanings are only in the head. Consciousness is the high-level pattern which brings the whole thing together, and emotional and moral self-control may well be a matter of how closely overall consciousness binds lower and more partial patterns of activity.

In Freeman’s view, the process of perception and action is not a two-way matter of inputs and outputs, but a one-way street of action on the world. Many people would agree that perception is an active business, not just a passive reception of impressions, but the idea that it consists entirely of action on the world sounds bonkers, and in fact Freeman does allow the outside world to influence our behaviour – the point is that all the ideas and interpretations bubble up from inside, and merely survive or fail to survive the impact of external reality. This is rather reminiscent of Edelman’s views and his analogy with the immune system, but Freeman draws from it the rather bleak conclusion that we are all, in a sense, in a state of solipsistic isolation from the world.

This creates a special problem for Freeman: how is it that we ever manage to overcome our isolation and communicate with each other? He sees social interaction as playing a mediating role, with processes rather similar to those which go on in the brain operating in the wider social

sphere - though not so similar that society itself becomes a conscious entity. Freeman has a number of ideas about signalling and communication to offer, but I'm not sure he really manages to deal with the underlying problem, and it remains a weak spot in the theory.

What, then is the answer on free will? At times Freeman seems to assert free will, while at others he seems to deny it: in fact he ultimately considers the question an ill-formed one. We see actions in terms of freedom or determinism because we are wedded to linear causality, even though we know that it does not provide an adequate view of the world, and that circular causality and more sophisticated perspectives are often more appropriate. For the swirling chaotic patterns of the brain, dynamic analysis is a more appropriate tool than those based on linear causation, and when we apply it correctly, the old opposition between free and determined is no longer an issue.

There's something in this, undoubtedly, but it doesn't dispel the sense of mystery which has made the old debate such a long-running philosophical staple. There does seem, intuitively at least, to be something uniquely odd about the causality of our minds, but if the problem arose entirely from a lack of dynamic analysis, we should surely find some of the causality of the normal world more mysterious than we do?

The Conscious Dead

10 July 2004

I don't know who first introduced them into philosophy, but zombies are frequently quoted in discussions of consciousness. Perhaps the obsession with brains which Hollywood zombies share with cognitive scientists has something to do with it.

Philosophical zombies are indistinguishable from normal people: their appearance and behaviour is perfectly normal, but they have no inner life: no phenomenal experience, no qualia, just colourless information processing. There isn't, to use Nagel's phrase, 'something it is like' to be a zombie. They are, to use a word which is no longer so unambiguous as it once was, robotic. Philosophical zombies are therefore, in fact, quite different from Haitian zombies or the Hollywood variety.

Philosophical zombies are used in various ways, but the main point about them is that they provide one of the chief arguments for the existence of qualia. The key question is: could zombies exist? According to one school of thought, it is intuitively obvious that they could, and that fact establishes that qualia are an important part of the riddle of consciousness - perhaps the most important part. The other main view is that zombies could not exist. You might take this view if you believed zombies would necessarily have to have qualia if they were to behave in exactly the way we do; or you might think that there are just no such things as qualia, anyway - so that in the relevant sense we are all effectively zombies anyway.

e

Zombies are often conceived of as being identical to their qualia-having equivalents, on an atom-by atom basis. Dennett 's zombies, by contrast, only resemble normal people externally (Dennett finds the whole zombie idea ridiculous, so the atom-by-atom version is probably just too extreme for him, even as a debating position). Inside, any kind of jiggery-pokery could be going on, so long as it does the (presumably computational) job required. He goes on, however, to propose a super-zombie or 'zimbo', which besides its routine processing of inputs and outputs, is capable of addressing and 'considering' its own internal states. He asks the interesting question, what would such a zimbo 'think' about its own experiences?

It would presumably think it had qualia. In fact, if all of the external behaviour of a zombie (or zimbo) exactly matches that of a normal human being, then a zombie would somehow have to talk as if it had qualia even though it didn't. It might well write a phenomenological dissertation on the subject, which would be indistinguishable from that of a real human writer. This is one of the main problems for a proponent of zombies. If they really are indistinguishable from normal human beings apart from qualia, that seems to make qualia a kind of ghostly irrelevance of a kind we'd be better off without.

It can't be denied, though, that the idea still seems plausible. Suppose we were cataloguing the badly-organised stock of a sweet-shop. While we note down the figures, our colleague calls out the quantities of red or black licorice found on various shelves. Even if we visualize red licorice the first time, after writing down the twenty or thirty figures, we are surely not having any red

qualia in association with the red licorice figures: but we are certainly acquiring information about the redness of various items. It seems perfectly reasonable to think that a zombie could function quite well with this kind of information, without ever having the kind we get when we actually see the licorice.

A second difficulty is over the issue of possibility. It seems unlikely on the face of it that zombies are possible in practice (Though how would we know? Everyone but you and I could be zombies – and I'm not completely sure about you), but it's generally felt that that isn't really necessary. They only have to be possible in theory – but what exactly does that mean?

Chalmers surely sets the bar too low when he claims that the mere intelligibility of the notion of zombies is enough. The idea of light without electromagnetic radiation is intelligible, but it does not establish that electromagnetic radiation is something over and above ordinary light. A more reasonable demand is that they should be logically possible – that is, they don't have to be compatible with the laws of physics, but they must not involve contradictory suppositions. On these terms, zombies seem acceptable, but it could be argued that that only establishes that qualia exist, or could exist, in some other world with different laws of physics, whereas we are really concerned with the world we actually live in.

Perhaps, then zombies have to be possible in this world – and why not? Well, it is a basic assumption we generally make that under the same conditions, the same events occur. This is a hard assumption to give up, because if different events could occur in identical circumstances, the world would become much harder to understand and predict, perhaps even entirely incoherent. But if philosophical zombies existed, we would have two identical physical sets of circumstances (one with me, one with my zombie twin) in which wholly different qualic events occurred.

In the final analysis, I think even the most committed zombists would accept that the argument is an appeal to our intuitions rather than a knock-down logical one. But the continuing interest in the issue of qualia shows how strong those intuitions are.

Feelings

20 July 2004


Recent theories of consciousness have tended to neglect or marginalise the emotions. I suppose this is because the dominant conception of thought these days sees it as a computational or at any rate, a problem-solving process. It isn't obvious where emotions fit into a program about high-level control of visual and motor systems. Perhaps there's also an unconscious feeling that emotions are wishy-washy stuff, unworthy of the attention of proper scientists. But they clearly have to be included in any full account of human mentality

With a few exceptions (such as Aaron Sloman's worthy attempts to give emotions a computational reading), most recent theories either disregard them or treat them as something bolted on to the main cognitive process. Perhaps they are the thing that provides the mental computer with motivation and goals? Perhaps they are the relics of older mammalian or even reptilian mental processes? Perhaps they are some kind of singalling process which plays a role in game theoretic transactions between human beings?

At any rate, they can (it's generally assumed) be left on one side while considering the purely rational functions of the mind, rather in the way that connectionist networks generally leave aside all the vastly complex chemistry of real biological neurons.

In 'The Feeling of What Happens' Antonio Damasio offers a very different account, one which integrates the emotions with both consciousness and the emergence of self and personhood. A holistic, or at least an integrated view is clearly something close to Damasio's heart: in 'Descartes Error ' he attacked the dualism which puts a gulf between the spirit and the body (or the mind and the brain, for that matter).


I really think poor Descartes deserves a bit of a break in this respect. OK, he was a dualist, but he wasn't the first dualist, nor the last. I think it's even debatable whether the net effect of his work was to reinforce or erode the position of dualism overall.


In any case, according to Damasio, consciousness, selfhood, and the emotions all spring ultimately from a single source: awareness of the current state of the body. The way they develop is complex, but this shared underlying focus neatly clarifies the relationship between them and ensures a close integration all the way up.

So far as the emotions are concerned, Damasio's idea resembles an earlier one put forward by William James. Putting it crudely, James said that it is the state of your guts which constitutes your emotions and influences your brain - not your brain which conceives emotions which then influence your guts. Damasio's theory gives a sophisticated development of this¶down-to-earth insight and lacks the debunking, reductive overtones of the earlier version. He draws a useful but novel distinction between emotions, which he defines as states of the body related to excitement or arousal, and feelings, which are the patterns evoked in the brain by emotions.

Many recent theorists have fought shy of the self or even dismissed it as an illusion, but Damasio proposes a whole hierarchy of selves, the lowest level of which is the proto-self. This is merely a short-term collection of neural patterns of activity which represent the current state of the organism. The core self is a second-order entity which maps the state of the proto-self in rather the same way the proto-self maps the current state of the body: whenever an encounter with an object impinges on the proto-self, the change is registered by activity in the core self. The core self represents the first, lowest level which deserves to be regarded as conscious, though this is the kind of immediate, unreflecting consciousness presumably possessed by animals in general, not just by human beings.

With the next step up, we get the autobiographical self, which draws on permanent (though modifiable) memories instead of just the immediate experiences which power the core self. At this point, there is a real, though still pre-linguistic, sense of self. Damasio thinks chimpanzees and probably dogs enjoy this level of consciousness. A final layer of development, with greater use of longer-term memory, delivers the kind of foresighted, reflective consciousness which we typically associate with human beings


That's rather a lot of different varieties of self for someone who's supposed to be a proponent of integration, isn't it? It's not even the full list, is it? We've also got the 'as if' body loop, which allows, as it were, hypothetical states of the body to be represented and considered - that seems to imply an 'as if' self. And then, on top of consciousness itself, we have conscience. Conscience seems a different kind of thing altogether to me - a function of consciousness, not a variety of it - but Damasio has it as the pinnacle of mental development.

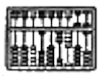

You normally like the moral element emphasised, don't you? You could say that the point about conscience was to bring the moral sense, too, within the same general framework. Again, I think it's rather neat to regard conscience as another higher-order level. Our normal desires are a matter of wanting more x, but morality is arguably a matter of desires about desires - we wish we didn't want certain kinds of x and resolve to over-ride our first-order desires by abstaining? That's not a point Damasio makes, but it does show how nicely his theory coheres, I think.

Anyway, all this is backed up by plausible neurological ideas, but the most convincing aspect is the way it ties emotional states of the body into the process of thought and reasoning. In Damasio's view, certain emotions (which in his account are states of the body, don't forget)get associated with certain possible contingencies or outcomes in the external world: they function as 'somatic markers' which allow us to be steered towards certain options, and react appropriately to potential longer term rewards, instead of responding to the immediate situation around us. Damasio backs this up by quoting examples of people with emotional impairment arising from specific brain lesions: as his theory predicts, they have trouble dealing effectively with the world even though their purely rational thinking is unaffected.


There's a problem with these examples, though, isn't there? If the emotional faculties can be knocked out by a few specific lesions in the brain, leaving the rational faculties still working, the two can't be as closely intertwined as all that, can they?

It's not clear to me, moreover, that the examples really are cases of people whose decision-making per se is impaired. We don't get a long account, but it sounds as if the problems are partly a matter of judging how to behave towards other people and partly a difficulty over managing one's own desires coherently. Those are bad problems to have, but they sound to me like the loss of particular mental modules rather than the impairment of central mental processes.

But my main problem with all this is that it skates over the real issues. Damasio talks about the proto-self representing the state of the body without any explanation of how merely being affected by something turns into being a representation of that something. He gives only cursory attention to the whole issue of qualia…


Well, you know, not every book has to be about your beloved qualia. And what Damasio does say (a brief diagnosis of the favourite story about Mary the colour scientist) is spot on, in my view. Understanding the process of having an experience is not the same as having the experience.

But I understand your attitude now. In this respect, and in his friendliness towards the 'multiple drafts' idea, Damasio is basically Dennettian (in other respects he certainly isn't, speaking with respect of both Searle and McGinn). Your problem is that Damasio demonstrates how a Dennett-style account is capable of being developed into an even more comprehensive theory, which is anathema to you, of course.

Francis Crick

9 August 2004

The death of Francis Crick at the end of last month drew many eulogies, most of which naturally highlighted the discovery of the double helix structure of DNA. In the latter part of his life, however, Crick turned his attention to the problem of consciousness rather than genetics or microbiology. He seems to have been temperamentally inclined to working in collaboration with a partner - having cracked the mystery of DNA with Watson, he now developed a fruitful partnership with Christof Koch. His view of consciousness, however, was summed up in his own book 'The Astonishing Hypothesis'.

The hypothesis in question is '…that "You", your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules.'

It has been suggested by some that this is not such an astonishing hypothesis after all. Certainly the idea that our mental life arises from the activity of the brain has been a mainstream one for a considerable time, but the key words in Crick's hypothesis are perhaps 'no more than'. On the face of it, this is indeed a fairly extreme claim; a reductionism which stops just short of denying the actual existence of consciousness. The quotation marks around the word 'You' suggest that Crick was also tempted by scepticism about the self.

It's difficult to be absolutely clear about Crick's philosophical position, however. Searle criticised 'The Astonishing Hypothesis' for not being clear about exactly what kind of reductionism it was putting forward, and with some justice: at times Crick talks in terms of emergence, and he seems to want to disavow naive or eliminativist reductionism, but his bottom line does seem to be that consciousness is nothing more than the activity of neurons.

One reason for this lack of clarity is perhaps that Crick, as he very fairly points out, is not even trying to set out a finished theory, only a hypothesis and a suggested line of attack. But the fundamental reason is that Crick is really interested in telling a scientific story, not a philosophical one. Most of the book is taken up with doing this. Crick's strategy is to approach consciousness via a consideration of the faculty of vision. He gives a very clear and interesting account of research in this area, with a well-judged balance of speculation and caution. Personally, however, I think the focus on vision dooms the enterprise from the start, at least as far as consciousness is concerned. The best one can hope to get by investigating vision alone is some insight into attention and sensory awareness; the central issues of consciousness are likely to remain untouched. Blind people are fully conscious, after all!

It's also true that Crick's close focus on neurons at the expense of philosophy seems to lead him into some dubious positions. He and Koch are particularly known for the view that consciousness arises when sets of neurons fire in a co-ordinated way, at frequencies around 40 Hertz. Crick suggests that synchronised firing of this kind might, in particular, be the neural correlate of visual awareness. To be really consistent with Crick's general attitude, the firing really needs to be visual awareness, not just correlated with it, but that is perhaps a nit-picking

point: the more fundamental difficulty is that no explanation is ever offered as to why co-ordinated firing should give rise to conscious experience. Crick suggests that this kind of co-ordination might be the answer to the notorious binding problem, because it explains how neurons in different visual areas which respond to different qualities of the seen object (form colour, motion, etc) 'temporarily become active as a unit', but it seems that at best that might be part of the answer. A particularly difficult aspect of the problem is that different pieces of sensory data which relate to the same object don't arrive in one place in the brain at the same time, yet our conscious experience never seems to suffer from, as it were, faulty lip sync. It's hard to see how simultaneous patterns of firing could deal with the chronological problem.

At the end of the book, Crick offers a short and tentative postscript setting out an idea about free will. This is really an explanation for why people think they have free will - Crick is presumably a determinist. His idea is that there is an unconscious part of the brain which makes the plans for what we are going to do: these plans then pop into the conscious mind as if from nowhere, giving an impression of free will. The conscious mind may be able to guess the factors behind the plans, or it may get them wrong: either way, it feels there is some mystery about the process. Crick, drawing on some research by Damasio, goes so far as to suggest that this unconscious planning facility (the 'seat of the will') is probably located in or near the anterior cingulate sulcus.

Of course it is perfectly true that the processes which give rise to conscious thought are not themselves conscious (otherwise we should be caught in a vicious regress), but that does not imply that consciousness is not in the driving seat. Often when we make a complex decision or draw up an explicit plan, we weigh the factors and consider possible events consciously in our minds, and it seems very hard to believe that this kind of process, which surely bears a remarkable resemblance to decision-making, is not ultimately responsible for the plan or decision which is eventually arrived at. Indeed, I think most people believe that making decisions and plans, and allowing human beings to rise above the influence of their immediate current environment, is exactly what consciousness is for.

As Crick himself would probably have been the first to agree, 'The Astonishing Hypothesis' is not the place to look for philosophy: but ten years after publication, in a fast moving field, it is still worth reading

Twin Earth

20 August 2004

Once upon a time (1750, in fact – or in any case, some time before the development of modern chemistry) there were two very similar planets. One was Earth: the other was also called Earth, but to save confusion let's call it Twin Earth. The general size, geography, climate and biology of Twin Earth were all pretty much like those of Earth – in fact, the two planets were virtually indistinguishable, with each person having an identical twin on the other planet.

However, there was one particular difference. On Twin Earth, the transparent liquid which made up the seas, lakes and rivers, which the animals drank, and which fell as rain - the substance, in fact, which the locals called 'water' – was not H2O, but XYZ. In most respects, and without resort to more sophisticated chemistry, it was impossible to spot any difference between the qualities and behaviour of XYZ and those of H2O.

(At this point, scientifically inclined readers may look worried and begin saying things like 'Yeah, but look... it couldn't be exactly the same as real water'. Well no, not really: with a bit of thought we could probably find a more plausible version, but H2O versus XYZ was what Hilary Putnam originally chose, and it doesn't really affect the argument.)

The strange result is that when Robinson, on Earth, thinks about the contents of his glass, he is thinking about H2O. But when Twin-Earth Robinson thinks an exactly similar thought, with exactly similar brain states, he is thinking about XYZ. The difference in what they are thinking about arises entirely from differences in the external world, not from any difference in the two brains. In the words of Putnam's famous slogan 'meaning isn't in the head'.

In one way, this doesn't seem so surprising: it seems almost common sense that meaning is affected by context. On the other hand, it seems a natural assumption that what you think about is pretty much under your own control, something you arrange for yourself within your own skull irrespective of the outside world. The Twin Earth argument undercuts the idea that meaning can arise from mental images or representations alone, which raises a difficulty for anyone wanting to endow a computer with consciousness, and anyone applying a functionalist interpretation to human consciousness. Computational representations in one's head cannot, it seems, be the same thing as psychological propositions in one's mind, at least not without some further ingredient.

The problem is illusory. Dennett sees this. First, he suggests, consider the behaviour of the coin-recognising mechanism in a slot machine. It may have been designed for American coins, but what if it is put to work 'recognising' Panamanian quarter-balboas (which are the same shape and weight). Do we say the machine is still (mistakenly) recognising US quarters, or has it somehow switched over to being a quarter-balboa recogniser? Who cares? If we like we can say that the intention of the designer means that the machine is still a US-quarter-recogniser, or equally, that the person who installed it in Panama has effectively transformed it into a 'q-balber' instead. In the end, we can interpret it whichever way suits us, can't we?

OK then. Now consider Twin Earth. Let's suppose, instead of the water case, we think about horses. On Twin Earth, let's say, they have schmorses instead; animals which resemble horses closely apart from genetics and some internal details. If some helpful philosophical aliens transport Robinson from Earth to Twin Earth, and he sees a schmorse, what does he mean when he says 'there's a horse'? Does he still mean 'horse', or does he now really mean 'schmorse'?

Isn't this the same as the coin machine? Horse, schmorse, what's the difference?


But the two cases aren't the same. It's OK to decide arbitrarily in the case of the coin machine, because the coin machine doesn't mean to identify any particular coin. It does whatever we intend it to do. But my thoughts don't depend on how you interpret my behaviour. The machine only has derived intentionality – any meanings come from the designer or user. But people have real, original intentionality. When I mean something, I mean it all by myself, irrespective of how other people may construe my meaning.


Ah, but that's just where you're wrong, and that's what Dennett explicitly denies. On your argument, meanings must remain forever a magic mystery: if you want a rational explanation, you have to accept that original, intrinsic meaningfulness is absurd. How can anything, even a brain, mean anything intrinsically?


I might ask, if all intentionality is derived, where does it ultimately come from? Dennett seems to think we can generate it from nothing by, as it were, taking in each other's washing. But if there's one thing that's clear to me, it is that the meaning of my thoughts doesn't in itself depend on other people's interpretations. I'll agree that Dennett is right about one thing though – this is one of the key issues, where people's intuitions divide sharply – almost as if they were on different planets

Cogito and Consciousness

30 August 2004

Poor Descartes, as I have remarked before, usually comes in for a ritual denunciation in books about consciousness, blamed for having espoused or even invented dualism, the doctrine of the separation of body and spirit. I think it would be more accurate to see him as presenting a new secular and rational perspective on the conception of the soul which was then prevalent (and to a considerable extent, still is); one which actually narrowed down its sphere of influence to the pineal gland. It seems perverse to me to suggest that later European thinkers got their dualism from Descartes rather than from the common Christian heritage, though that seems to be the way most people see it

But Descartes is relevant to the current debate about consciousness in other ways, too. In particular, I think his most famous argument, the 'cogito', challenges some currently popular perspectives in a way which is well worth considering.

The cogito ('cogito ergo sum' - I think, therefore I am) is surely the most well-known argument in philosophy - it occupies the kind of place in its field which the Mona Lisa, Hamlet's soliloquy, or Beethoven's Fifth Symphony, occupy in theirs. This kind of mass popularity can, paradoxically become a barrier to proper appreciation. In order to draw out its importance for consciousness, I should like to focus on two points about it: first, it isn't original; second, it isn't logical.

Not original, because the same argument can actually be found in St Augustine. I don't think anyone knows for sure whether Descartes found it there or came up with it independently, but the obvious question is - if St Augustine came up with it first, why aren't we all talking about St Augustine's cogito?

The reason, I think, is that the argument was of no great importance to St Augustine. He was in pursuit of faith, not doubt, and was more interested in God's existence than his own. He puts no particular stress on the cogito argument, and readers who aren't particularly looking out for it could easily read it without noticing its significance. For Descartes, by contrast, everything depended on it. He needed a point of certainty from which to begin the construction of his metaphysics: St Augustine already had a source of certainty in God. Descartes also turned to God as a guarantor of knowledge, of course, but in his case, unprecedentedly, God did not come first. In this respect, modern philosophers are mostly in the same boat as Descartes, and if they want certainty, they have to undertake a similar exploration.

Not logical? I don't, of course, mean that the cogito is illogical, or contains flawed reasoning. But the cogito is often interpreted as a piece of pure logic, an a priori argument whose truth does not depend on observation, like the truths of mathematics. This is a mistake.

Suppose the characters '2 + 2 = 4' were to appear by chance in the pattern of clouds in the sky: the statement they represent would still be true. But if the clouds somehow lined up to form the words 'cogito ergo sum', it would be false - the clouds are not doing any thinking. Descartes is not offering a logical proof, but making a claim that certain kinds of perception, or thought, are

immune from error. In particular, when I perceive my own existence, I cannot be wrong, because if I didn't exist no perceiving would be going on. Even doubting my own existence actually proves it, because only entities which exist can doubt their own reality.

There may be just a few other perceptions which have a similar immunity from error. Arguably, you can't be wrong about being in pain, for example, though you might be wrong about the existence of the dentist. But I digress...

The point here is that if you're looking for philosophical certainty, and you can't get it from God, it can only be found in the first-person view, and in the self. But the first-person view, and the reality of the self, are just what many modern thinkers about consciousness would want to do without. These days, we look to empirical science for the truth, and distrust our own inner phenomenal experience, although all our knowledge of the world, and of science, actually derives in the final analysis from interpretation of our own subjective phenomenal experiences (doesn't it?).

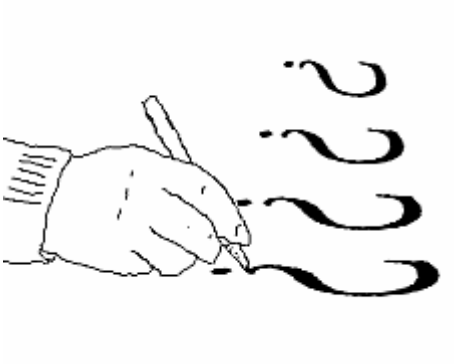
It wasn't always quite like this. In Brentano's day it was natural to think that the business of psychology included the classification and study of one's own inner, subjective sensations: but the disastrous over-development and subsequent collapse of introspectionist psychology functioned like a nuclear explosion, not merely destroying the existing structures, but rendering the whole territory of phenomenal experience uninhabitable and even unvisitable for a generation. During the era of Behaviourism, subjective experience and even consciousness itself was actually denied as a result. Those days have passed, but many still feel that if something can't be investigated from the third-person point of view, it can't be brought within respectable science at all.

I suppose this epitomises the dilemma of consciousness as it exists today: direct, first-person investigation of phenomenal experience seems to lead nowhere, but certain key aspects of consciousness - qualia, meaning, selfhood, and so on - seem to have no place in the objective third-person account.

Some, of course, are bold enough to offer eliminative accounts of these intractable phenomena: for them, the Cartesian suggestion that all knowledge ultimately springs from phenomenal experience must surely be profoundly unpalatable. Could this be one of the reasons why Descartes keeps getting such a drubbing?'

Knowing

10 September 2004



Carl Zimmer described some interesting research in a recent blog entry. It seems that people who are unable to recall any of the events of their past lives are still able to identify which of a list of words best describes them as people: although their explicit knowledge of their own autobiographies has disappeared, they still have self-knowledge in a different form. It is suggested that two different brain systems are involved. This research might possibly shed a chink of light on the debate about whether, and in what sense, we actually have selves, but it also raises the thorny question of different ways of knowing things. We often talk about knowing things as though knowledge was a straightforward phenomenon, but it actually covers a range of different abilities – look at the following examples'

- I know what the capital of Ecuador is.
- I know where the keys are.
- I know how to sign my name.
- I know that zebras don't wear waistcoats

The first example is the case in which an explicit fact has been memorised – possibly even a fixed formula ("The capital of Ecuador is Quito."). This is perhaps the easiest form of knowledge to deal with in a computational way – we just have to ensure that the relevant string of characters ("Quito") or the appropriate digits are stored in a suitable location. There's relatively little mystery about how you can articulate this kind of knowledge, since it has probably been saved in an articulated form already: it was words on the way in, so it's no surprise that we can provide words when it's on the way out.

In the second case, things are slightly less clear. It's unlikely, unless you have a really bad key-losing problem, that you have memorised an explicit description of the place where they are: however, if you need to produce such a description, you would normally have no particular difficulty in doing so. A reasonable assumption here might be that the relevant data on the position of the keys are still stored somewhere explicitly, (on some sort of map, as co-ordinates, or perhaps more likely, as a set of instructions like those on pirate's treasure maps, telling you how to get to the treasure/keys). The question of how you are able to translate this inner data into a verbal description when you need one is less easily answered, but then, the process of coming up with a description of anything is not exactly well understood, either.

The third case is a bit different. The importance of 'knowing how' as a form of knowledge was emphasised by Gilbert Ryle as part of his efforts to debunk the 'Ghost in the Machine'. It could legitimately be argued that the difference between 'knowing how' and 'knowing that' is so great that it makes no sense to consider them together - but both do allow some stored knowledge to influence current behaviour, so there is at least that broad similarity. You sign your name without hesitation, but describing how to do it (unless you happen to be called 'O') is challenging. To describe the required series of upstrokes and downstrokes would require careful thought – you might even have to watch yourself signing and take notes. The relevant data must be in your brain or your hand somewhere, but you are hardly any better off when it comes to putting them into words than anyone else who happens to be watching. Presumably the

relevant data are still stored somewhere in your brain. Perhaps they are just in a different part of it, or otherwise less accessible: but it seems likely that at least some of them are held in a form which just doesn't translate into explicit terms. There may be a sequence of impulses recorded somewhere which, when sent down the nerves in your arm, results in a signature: but there need be no standard pattern in the sequence which symbolises 'downstroke' or anything else.

The fourth case is the most difficult of all. Most of the things we know, we never think about or use. We never asked ourselves whether zebras wore waistcoats until Dennett proposed the example, but as soon as we heard the question, we knew the answer. This vast stock of common-sense knowledge (Searle refers to it, or something very like it, as 'the Background') is crucial to the way we deal with real life and work out what people are talking about: it's the reason human beings don't generally get floored by the 'frame problem' – unanticipated implications of every action - the way robots do. It surely cannot be that all this kind of knowledge is saved explicitly somewhere in the brain, however vast its storage capacity. In fact, there are good arguments to suggest that the amount of information involved is strictly infinite: we know zebras are normally less than twenty feet tall, normally less than twenty-one feet tall, and so on.

That last argument suggests a better explanation – perhaps key pieces of information are stored in a central encyclopaedia, and the more recondite conclusions worked out as necessary. After all, if we know that zebras are less than twenty feet, simple arithmetic will tell us that they are less than fifty, without the need to store that conclusion separately. The trouble then is that there simply is no general method of working out relevant conclusions from other facts: formal logic certainly isn't up to the job. I've discussed this further elsewhere, but it seems likely to me that part of the problem is that, as with the third case, the information is probably not recorded in the brain in any explicit form. To look for an old-fashioned algorithm is probably, therefore, to set off up a blind alley.

What about the two forms of self-knowledge we started with? It looks as if we are dealing with a loss of the kind of knowledge covered by example 1 above, while type 2 is retained. If so, the loss of memory might be less disabling than it seems. The patients in question would not be able to tell you their address, but perhaps they might still be able to walk to the correct house "without thinking about it"? They might not be able to tell you their own name, even - but perhaps they could sign it and then read it.

Robot Ethics

20 September 2004

 The protests of Isaac Asimov fans about the recent film I Robot don't seem to have had much impact, I'm afraid. Asimov's original collection of short stories aimed to provide an altogether more sophisticated and positive angle on robots, in contrast to the science fiction cliché which has them rebelling against human beings and attempting to take over the world. The film, by contrast, apparently embodies this cliché. The screenplay was originally developed from a story entirely unrelated to I Robot : only at a late stage were the title and a few other superficial elements from Asimov's stories added to it.

As you probably know, Asimov's robots all had three basic laws built into them:¶

- A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The interplay between these laws in a variety of problematic situations generated the plots, which typically (in the short stories at least) posed a problem whose solution provided the punchline of the story.

 I enjoyed the stories myself, but the laws do raise a few problems. They obviously involve a very high level of cognitive function, and it is rather difficult to imagine a robot clever enough to understand the laws properly but not too sophisticated to be rigorously bound by them: there is plenty of scope within them for rationalising almost any behaviour ("Ah, the quality of life these humans enjoy is pretty poor - I honestly think most of them would suffer less harm if they died painlessly now.") It's a little alarming that the laws give the robot's own judgement of what might be harmful precedence over its duty to obey human beings. Any smokers would presumably have the cigarette torn from their lips whenever robots were about. The intention was clearly to emphasise the essentially benign and harmless nature of the robots, but the effect is actually to offer would-be murderers an opportunity ("Now, Robbie, Mr Smith needs this small lead slug injected into his brain, but he's a bit nervy about it. Would you...?"). In fairness, these are not totally dissimilar to the problems Asimov's stories dealt with. And after all, reducing a race's entire ethical code to three laws is rather a challenge – even God allowed himself ten!

The wider question of robot ethics is a large and only partially explored subject. We might well ask on what terms, if any, robots enter the moral universe at all. There are two main angles to this: are they moral subjects, and if not, are they nevertheless moral objects? To be a moral subject is, if you like, to count as a person for ethical purposes: as a subject you can have rights and duties and be responsible for your actions. If were is such a thing as free will, you would probably have that, too. It seems pretty clear that ordinary machines, and unsophisticated

robots which merely respond to remote control, are not moral subjects because they are merely the tools of whoever controls or uses them. This probably goes for hard-programmed robots of the old school, too. If some person or team of persons has programmed your every move, and carefully considered what action you should output for each sensory input, then you really seem to be morally equivalent to the remote-control robot: you're just on a slightly longer lead.

 Isn't that a bit too sweeping? Although the aim of every programmer is to make the program behave in a specified way, there can't be many programs of any complexity which did not at some stage spring at least a small surprise on their creators. We need not be talking about errors, either: it seems easy enough to imagine that a robot might be equipped with a structure of routines and functions which were all clearly understood on their own, but whose interaction with each other, and with the environment, was unforeseen and perhaps even unforeseeable. It's arguable that human beings have downloaded a great deal of their standard behaviour, and even memory, into the environment around them, relying on the action-related properties or affordances of the objects they encounter to prompt appropriate action. To a man with a hammer, as the saying goes, everything looks like a nail: maybe when a robot encounters a tool for the first time, it will develop behaviour which was never covered explicitly in its programming.

But we don't have to rely on that kind of reasoning to make a case for the agency of robots, because we can also build into them elements which are not directly programmed at all. Connectionist approaches leave the robot brain to wire itself up in ways which are not only unforeseen, but often incomprehensible to direct examination. Such robots may need a carefully designed learning environment to guide them in the right directions, but after all, so do we in our early years. Alan Turing himself seems to have thought that human-level intelligence might require a robot which began with the capacities of a baby, and was gradually educated.

 But does unpredictable behaviour by itself imply moral responsibility? Lunatics behave in a highly unpredictable way, and are generally judged not to be responsible for their actions on those very grounds. Surely the robot has to show some qualities of rationality to be accounted a moral subject?

 Granted, but why shouldn't it? All that's required is that its actions show a coherent pattern of motivation.

 Any pattern of behaviour can be interpreted as motivated by some set of motives. What matters is whether the robot understands what it's doing and why. You've shown no real reason to think it can.

 And you've shown no reason to suppose it can't.

 Once again we reach an impasse. Alright, well let's consider whether a robot could be a moral object. In a way this is less demanding – most people would probably agree that

animals are generally moral objects without being moral subjects. They have no duties or real responsibility for their actions, but they can suffer pain, mistreatment and other moral wrongs, which is the essence of being a moral object. The key point here is surely whether a robot really feels anything, and on the face of it that seems very unlikely. If you equipped a robot with a pain system, it would surely just be a system to make it behave 'as if' it felt pain – no more effective in terms of real pain than painting the word 'ouch' on a speech balloon.


Well, why do people feel pain? Because nerve impulses impinge in a certain way on processes in the brain. Sensory inputs from a robot's body could impinge in just the same sort of way on equivalent processes in their central computer – why not? You accept that animals feel pain, not because you can prove it directly, but because animals seem to work in the same way as human beings. Why can't that logic be applied to a robot with the right kinds of structure.


Because I know – from inside – that the pain I feel is not just a functional aspect of certain processes. It actually hurts! I'm willing to believe the same of animals that resemble me, but as the resemblance gets more distant, I believe it less: and robots are very distant indeed.


Well, look, the last thing I want is another qualia argument. So let me challenge your original assumption. The key point isn't whether the robot feels anything. Suppose someone were to destroy the Mona Lisa. Wouldn't that be a morally dreadful act, even if they were somehow legally entitled to do so? Or suppose they destroyed a wonderful and irreplaceable book? How much more dreadful to destroy the subtle mechanism and vast content of a human brain – or a similarly complex robot?


So let me get this right. You're now arguing that paintings are moral objects?


Why not? Not in the same way or to the same degree as a person, but somewhere, ultimately, on the same spectrum.

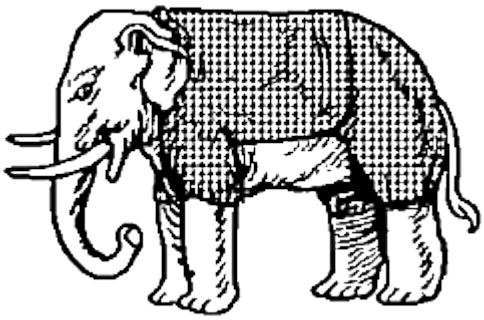

That's so mad I don't think it deserves, as Jane Austen said, the compliment of rational opposition.

Implicature

1 October 2004

“I’m leaving”

“Who is she?”



This is the kind of human exchange which computers, it is said, will never be able to fathom. The computer might be able to decode the literal meaning of the sentences, but never pick up the “relationship crisis” scenario which human beings recognise more or less instantly. This is an example of the exciting subject of conversational implicature, an ugly word invented by H.P.Grice to describe the inferences we make which are vital to understanding each other. Grice proposed that in order to work out what other people are getting at, we take it for granted that when they talk to us they are normally going to follow certain rules, or maxims. There are nine altogether, grouped under four headings:

Quantity

i) Make your contribution as informative as is required for the current purposes of the exchange.
ii) ii) Do not make your contribution more informative than is required.

Quality

i) Do not say what you believe to be false.
ii) ii) Do not say that for which you lack evidence.

Relational

Be relevant

Manner

i) Avoid obscurity of expression
ii) Avoid ambiguity
iii) Be brief
iv) Be orderly

These maxims help us express ourselves briefly without fear of being misunderstood. If someone tells you your horse came in either first or fourth, for example, you are entitled to assume, without his saying so, that he does not know which it was (unless other evidence suggests he is deliberately being annoying). Or consider this exchange:

A: Smith doesn’t seem to have a girlfriend these days.

B: He has been paying a lot of visits to New York recently.

If B is following the rules, we have to assume that what he says is relevant in some way, so we can legitimately deduce that he means Smith may have a girlfriend in New York. Besides straightforward use like this, the rules can be deliberately broken as a more subtle way of sending messages. If you provide a job reference which merely says that X worked for you and always turned up on time, the recipient may reasonably assume that you are deliberately being

uninformative as a way of implying a negative judgement about X which you prefer not to spell out explicitly.

Grice's work was directed towards philosophical ends, but it has proved unexpectedly fertile in other areas. In particular, it offers a promising-looking angle on the old debate about whether you can get syntax from semantics: maybe you don't need to but can look instead to the new field of pragmatics, in which the linguistic consequences of Grice's original insights are explored. Particularly interesting is Dan Sperber and Deirdre Wilson's book tackling the mystery of Relevance, a subject which (according to me, at least) is one of the three-and-a-half big problems of consciousness.

Sperber and Wilson describe two models of communication. In the code model, a thought in one brain (perhaps in mentalese, a brain's private mental language) is encoded into language, transmitted to another brain, and decoded again. In the inferential model, by contrast, the speaker just provides the linguistic evidence from which the auditor, in a more or less Gricean way, can infer the intended message. The code model cannot provide a complete explanation of the process of communication because it can't deal with the sort of example quoted above: on the other hand, when you examine real conversations, there does seem to be a whole lot of coding going on. Sperber and Wilson maintain that both code and inference have roles to play.

It's rather as though human communication had started out as a game of charades (the game where one player has to mime a book, play or film title for the others to guess). At first, the players have to think of clues that straightforwardly make the audience think along the right lines - behaving like a whale for Moby Dick, say. But experienced players gradually develop a set of conventions, such as pointing to their nose (="knows") with one finger and at someone who has guessed correctly with another. The grammatical and lexical apparatus of ordinary language represents the code, the ultimate development of this kind of helpful convention, and in ordinary communication it now does most of the work, though successful communication still requires occasional resort to inferential methods.

Sperber and Wilson think Grice's approach can be slimmed down, and most of the work done by the simple criterion of relevance. This means that the inferential component of communication rests very largely on the ability to work out what is and isn't relevant in what people are saying to you. They offer an analysis of relevance in terms of contextual effect. Roughly speaking, the idea is that a new piece of information is relevant to the extent that it allows you to revise the beliefs you had already, strengthening, weakening or deleting them or allowing new deductions to be added. If it merely duplicates things you already know, or if it has no connection with any of the pieces of information you have already, it isn't relevant at all. This, in a more formal guise, is how Sperber and Wilson propose the relevance of any given proposition could in principle be rated.

So, on their theory, if someone says "Coffee would keep me awake.", I am entitled to assume that the utterance has some relevance to something. As a first shot, I try out the simplest hypothesis I can think of - maybe they just think I need to know about the stimulating properties of coffee? But on that interpretation, the relevance of the remark seems negligible - we're not in a pharmacological seminar, and at best the remark might convey to me something I didn't know about coffee - it doesn't change any of my other current beliefs or allow me to draw many new conclusions. So I move on to more complex hypotheses, and eventually (rather quickly, I hope) I reach the correct hypothesis: depending on the circumstances, coffee is either being asked for or refused.

This looks like a promising avenue to explore, but there are some ifs and buts. First, the analysis is all to do with conversations, and whether it can be widened into any general analysis of relevance is unclear. Second, it's not at all clear how this approach could be reduced to the kind of concrete algorithm you might use in programming a chat-bot or other program. In particular, there seems to be a danger of circularity. In many cases, when we come to tot up the contextual effects of a given remark, there are actually going to be an infinite number of valid new inferences we can make -nearly all of them will be trivial or uninteresting, but how do we weed them out. I have the uneasy feeling that at some stage we are going to find ourselves needing to disallow inferences which are (gulp) irrelevant...

Even if that difficulty is a real one, the idea of implicature seems an interesting one. Sperber and Wilson are good at coming up with little snatches of dialogue which exemplify different varieties of implicature, and it is noticeable that many of them have a witty or rhetorical quality. Towards the end of the book, they discuss the way implicatures contribute to poetry, metaphor and other special forms of language. It seems to me that virtually all jokes are based on implicature, too. Consider this example:

G: Last night I shot an elephant in my pyjamas. How he got into my pyjamas I'll never know.

Surely a fine example of the manipulation of Gricean conversational implicatures..?',

Consciousness and Relativity

11 October 2004

Roger Penrose and Stuart Hameroff have famously suggested that consciousness is inextricably connected with quantum effects (albeit ones which we don't yet understand). Richard M Pico, less famously so far, has invoked the other great pillar of modern physics: relativity.

What on earth has relativity got to do with it? If we had to sum up the theory set out in Pico's book "Consciousness in Four Dimensions", we might choose the slogan 'life is a frame of reference'. Pico is not unique in emphasising the temporally extended nature of life and consciousness - Steven Rose, for example, has put forward a more general version of that perspective, and Locke's view that 'Nothing but consciousness can unite remote existences into the same person' seems very close to Pico's central insight.

Pico describes the emergence of life in a story of how protocells may have turned into true life. Protocells, on this view, have a simple bubble-like wall inside which certain reactions go on. But the wall is not enough to defend them from changes in the environment; when the right circumstances come along they form, and when the physical or chemical environment, in one of its periodic oscillations, becomes unfavourable, they simply fall apart again. Eventually a lucky protocell comes up with internal reactions which, fortuitously, have a homeostatic effect - they regulate the internal environment of the cell, defending it from external changes to the point where the cell can survive through the regular cycles of change in its immediate environment.

This persistence, on Pico's view, is what characterises life: more debatably, he characterises it as a 'frame of reference'. It certainly establishes a kind of physico-chemical baseline within the cell, but that seems to bear only a metaphorical relation to the 'frames of reference' proposed by relativity. A frame of reference (if I've understood correctly) is a point of view from which a particular set of measurements of the world and a particular perception of the simultaneity of events holds good: but the view from inside the cell is surely much the same as the view from outside. You might, I suppose, say that time passes differently within the cell because the normal external oscillation, which in a sense beats out time, has been dampened or stopped: but that would be a loosely metaphorical version of relativity.

Be that as it may, Pico sees the emergence of consciousness as broadly recapitulating the emergence of life. The neurological details of the theory are set out with admirable clarity, and with a level of detail it is impossible to do full justice to here. Briefly, Pico believes the columnar structures of the prefrontal neocortex provide prefrontal integration modules (PIMs) which bring together a wide and disparate range of sensory inputs. Generally, the patterns formed are transient, each being swept away by a succeeding wave of inputs: but in the course of evolution some of these PIMs acquire new properties in more or less the way the true cells raised themselves above the level of the protocells. They become able to retain an echo of previous states, and hence provide the basis for true perception of time, and consciousness.

This idea has some appeal. It is a common insight that while animal behaviour is generally governed by conditions in the present moment, conscious thought allows us to address goals in the remote future and adopt chronologically extended plans. It's also true that if we want to include a Self in our theory, it somehow has to have a continued existence over the full period of

the individual's life. As a third bonus, it allows Pico a neat view about the ontological reality of consciousness, namely, that it is as real as life (and we might add, as elusive).

However, there are two big problems. First, as with life, this doesn't really look like relativity in any but the loosest of senses. Second, and much worse, it just doesn't seem to explain the nature of consciousness. All it tells us is that consciousness has homeostatic properties, and retains items from its past: but you could say the same about parts of the digestive system

It is a good and valid point, however, that a lot of scientific thought outside the confines of physics itself still operates on a Newtonian basis. At the back of most our minds is the Laplacean idea that if we could specify all the data about every particle in the Universe at a single instant, the whole future and past would be calculable. This way of thinking, I believe, lies behind the intuitive certainty which some people feel that consciousness must ultimately be a computational phenomenon. After all, if the Universe itself is essentially a discrete-state machine, with one state-of-affairs arising directly from the preceding state-of-affairs, then everything must be computable. But that begs the question; and in fact, relativity denies the possibility, even in principle, of an objectively correct time-slice containing a full description of every point in the Cosmos at a single given moment. There is no such thing as absolute simultaneity, and if we want our determinism to be computable, we really need a more sophisticated version.

All in all then, I think you could say that Pico is on to something; unfortunately the thing he's on to is not, in the end, The Answer.

Medieval Chat-Bot

23 October 2004

Machines that deal with numbers and perform useful calculations have a long history, gradually increasing in power and flexibility over the course of several centuries. Machines which deal intelligently with words, and produce sensible prose, however, seem like a relatively recent aspiration. There were simple humanoid automata in Descartes' day, and impressively sophisticated ones during the eighteenth century: such 'robots' naturally gave rise to the speculation that they might one day speak as well as mimic human beings in other ways. But surely Turing was the first person to propose in earnest a machine which could produce worthwhile words of its own?

There were, of course, many more or less mechanical ancient systems designed to produce oracles or mystical insights. The I Ching is one interesting example: the basic apparatus is a collection of texts, with the appropriate one for a given occasion to be looked up using randomly-generated patterns. (the patterns are produced by throwing sticks, though coins and other methods can be used). If that was all there was to it, it would seem to be more or less computational in nature, albeit very simple – not very different, in some ways, to the sortes Virgilianae, the Roman system in which a random text from Virgil was taken as oracular. Leibniz took the symbols which identify the I Ching texts, which consist of sets of broken and unbroken lines, to be a binary numbering system, which would strengthen the resemblance to a modern program. In fact, although the symbols do lend themselves to a binary interpretation, that isn't the way they were designed or understood by the original practitioners. More fundamental, the significance of the results properly requires meditation and interpretation; it isn't really a purely mechanical business.

It is just conceivable (I think) that Roger Bacon considered the possibility of a talking machine. There is an absurd story of how he and another friar constructed a brass head which would have pronounced words of oracular wisdom had not their servant botched the experiment by ignoring the vital moment when the head first spoke. This tale was perhaps influenced by earlier stories about mechanical talking heads, such as the bull's head Pope Silvius (an innovative mathematician, interestingly enough) was said to have had, which would return infallible yes or no answers to any question. There is no evidence that Bacon ever contemplated a talking machine, but a procedure for generating intelligible sentences would have been the sort of thing which might have interested him, and I like to think that the brass head story is a distorted echo of some long-lost project along these lines.

In any case, there is one incontrovertible example from about the same time; Ramon Llull's Ars Magna provides a mechanical process for generating true statements and even proofs. Llull, born around 1232, enjoyed a tremendous reputation in his day, and was famous enough to have had his name anglicised as 'Raymond Lully'. He was better acquainted with Jewish and Arabic learning than most medieval scholars, and may possibly have been influenced by the Kabbalistic tradition of generating new insights through new permutations of the words and letters of holy texts. He wrote in both Arabic and the vernacular, and among other achievements is regarded as a founding father of Catalan literature.

The Ars Magna has a relatively complex apparatus and uses single words rather than extended texts as its basic unit. The core of the whole thing is the table below.

	Absolute principles	Relative principles	Questions	Subjects	Virtues	Vices
B	Goodness	Difference	Whether	God	Justice	Avarice
C	Greatness	Concord	What	Angel	Prudence	Gluttony
D	Eternity	Opposition	Whence	Heaven	Fortitude	Luxury
E	Power	Priority	Which	Man	Temperance	Pride
F	Wisdom	Centrality	How many	Imaginative	Faith	Sloth
G	Will	Finality	What kind	Sensitive	Hope	Envy
H	virtue	Majority	When	Vegetative	Charity	Anger
I	Truth	Equality	Where	Elemental	Patience	Untruthfulness
K	Glory	Minority	How - with what	Instrumental	Piety	Inconstancy

Llull provides four figures in the form of circular or tabular diagrams which recombine the elements of this table in different ways. Very briefly, and according to my shaky understanding, the first figure produces combinations of absolute principles – 'Wisdom is Power', say. The second figure applies the relative principles – 'Angels are different from elements'. The third brings in the questions – 'Where is virtue final?'. The fourth figure is perhaps the most exciting: it take the form of a circular table which is included in the book as a paper wheel which can be rotated to read off results. This extraordinary innovation has led to Llull acquiring yet another group of fans – this is regarded as the forerunner of pop-up books. The fourth figure combines the contents of four different table cells at once to generate complex propositions and questions.

The kind of thinking going on here is not, it seems to me, all that different from what goes into the creation of simple sentence-generating programs today, and it represents a remarkable intellectual feat. But the weaknesses of the system are obvious. First, the apparatus is capable of generating propositions which are false or heretical, or (perhaps more worrying to us) highly opaque, open to various different interpretations. Llull implicitly recognises this: in fact, to perform some of the tasks he proposes – such as the construction of syllogisms – a good deal of interpretation and reasoning outside the system is required: the four figures alone merely give you a start. Second, the Ars Magna is quite narrowly limited in scope – it really only deals with a restricted range of theological and metaphysical issues. Of course, this reflects Llull's preoccupations, but he also remained unworried by it because of two beliefs which then seemed natural but now are virtually untenable. One is that the truths of Christianity are ultimately as provable as the truths of geometry. The other is that the world is fully categorisable.

Llull's system, and many others since, is ultimately combinatorial. It simply puts together combinations of elements. In order to deal with the world at large successfully, such a system has to have a set of elements which in some sense exhaust the universe – which cover everything there is or could be. When we put it like that, it seems obvious that there is no way of categorising or analysing the world into a finite set of elements, but the belief is a tenacious one. Even now, some are tempted to believe that with a big enough encyclopaedia, a computer will be able to deal with raw reality. But any such project sends the finite out to do battle with the infinite. Perhaps it would help to realise just how old – how medieval – this particular project really is.

Gavagai

31 October 2004

Walking along one day on the newly-discovered coast of Australia, Captain Cook saw an extraordinary animal leaping through the bush.

"What's that?" he asked one of the aborigines accompanying him.
"Uh - gangurru." he replied - or something like that. Captain Cook duly noted down the name of the peculiar beast as 'Kangaroo'.
Some time later, Cook had the opportunity to compare notes with Captain King, and mentioned the kangaroo.
"No, no, Cook", said King, "the word for that animal is 'meenuah' - I've checked it carefully.
"So what does 'kangaroo' mean?"
"Well, I think," said King "it probably means something like 'I don't know'..."

But it was too late, and so ever since, the English word 'kangaroo' has been based on a misunderstanding, and really means 'I don't know'. At any rate, that's how the story (probably apocryphal) goes.

It may well have been this story which Quine had in mind when he came up with the 'Gavagai' example used in 'Word and Object'. The word has achieved a kind of fame in itself: there is now a species of beetle named after it, and it was used by Umberto Eco as the name of a minor character in his novel 'Baudolino'.

Quine was concerned to give a fundamental explanation of how we manage to use words: 'how surface irritations generate, through language, one's knowledge of the world'. He addressed in particular the problem of radical translation - how we can find a translation for words in an entirely unknown language which has no known correspondence with our own. Utterances such as 'ouch', he suggested, are relatively straightforward - the word properly corresponds with the occurence of pain. But other words, even nouns, do not correspond so exactly to the pattern of stimuli we happen to be experiencing at the time.

Suppose we have a linguistic explorer and a native subject: a rabbit runs past and the native exclaims 'Gavagai!'. The linguist forms the reasonable hypothesis that 'Gavagai' means 'rabbit', but how can he be sure? Putting aside some difficulties about 'yes' and 'no', he can ask the native in a series of different situations the simple question 'Gavagai?' and see what responses he gets.

Quine wants to build up from 'stimulus meaning', which is equivalent to a list of all the stimuli which would prompt the native to say 'yes' in response to the stereotyped question, to more complex and natural kinds of meaning - 'occasion sentences' which are true in particular sets of immediate circumstances (eg when a rabbit is present) and 'standing sentences' which are true irrespective of what happens to be going on around us at the moment, for example.

But the upward path proves unexpectedly difficult. As a matter of fact, the native's tendency to agree may be influenced by extraneous factors - he may have seen a rabbit a few minutes before

and hence be prepared to accept a mere rustling in the grass as sufficient evidence. Or he may observe a characteristic ‘rabbit fly’ unknown to the linguist which betokens the presence of a rabbit even in the absence of any other evidence. Or perhaps he dissents, not because he thinks there isn’t a rabbit, but because he presumes the linguist to be a hunter and there isn’t at the moment a clear shot available.

But even if we can get round these difficulties, there simply is no guarantee that any finite set of observations pin down the correct meaning of the word ‘gavagai’. Even if we can establish identity if stimulus meanings, we cannot thereby verify ‘intrasubjective synonymy’ of ‘gavagai’ and ‘rabbit’. The native may use the word in exactly those situations in which the linguist would use the word ‘rabbit’, but it could still mean something different: ‘temporal section of a rabbit’ or ‘set of undetached rabbit parts’. For that matter, it could mean ‘rabbit or dalek’ or ‘rabbit before the year 3000 and bear after that’. Ultimately Quine affirms the ‘indeterminacy of translation’ - we cannot provide certain radical translations (and since translation is, for Quine, much the same as understanding, we can never be sure we have correctly understood any linguistic utterance either).

A natural reaction to this disastrous conclusion is incredulity. Yes, of course, some degree of uncertainty attaches to everything; yes, we can always come up with a fantastic alternative meaning for any sentence which is logically consistent with the circumstances, but so what? We just know that in practice translation is possible. There are two things which can be said in reply.

First, problems of translation and misunderstanding are not as exceptional as all that. It would be entirely possible for an Englishman and an American to have lengthy conversations about robins without realising that the word refers to entirely different birds on different sides of the Atlantic (hence, in part, the bemusement which generally creeps over a British audience during one particular scene in ‘Mary Poppins’); or indeed, to talk about turtles without realising that one used the word in a significantly more inclusive sense than the other.

But second, Quine never denied that in ordinary conversation and translation we seem to manage pretty well - the question is how, exactly? It’s not difficult to explain that once you have seen enough examples of the use of ‘gavagai’ the meaning becomes obvious - but stating exactly how it becomes obvious, and how all those implausible alternatives are recognised as implausible, is extremely difficult, as the creators of translation programs have found. One can liken Quine’s attitude to translation to Hume’s puzzlement over induction - it seems to work: indeed, without it we should be in grave difficulty, but why it works is an utter mystery.

To quote one last practical example, we can return to poor old Captain Cook. Although the legend of his misunderstanding still thrives, it appears from more recent research that in fact the word he heard all those years ago probably did mean ‘kangaroo’ in a particular local dialect after all. So what about Captain King’s ‘meenuah’? It turns out that the word he was reporting was probably one that meant, not ‘kangaroo’, but ‘edible animal.

Konsciousness

15 November 2004


OK, a modest proposal from me. One of the big problems in sorting out consciousness is simply defining what it is (as you can see from this selection of views). Ned Block is absolutely right in calling it a 'mongrel' concept - a mixed-up mess of different ideas. He makes a distinction between access or a-consciousness, which has to do with thinking about plans and actually doing stuff, and phenomenal, or p-consciousness, which is to do with having experiences, 'raw feels' and all that wishy-washy qualia stuff. That is a valid and important distinction, but there are plenty of others. For some people, for example, consciousness is basically awareness - you can't be conscious without being conscious of something. Others see nothing absurd about being conscious without having any sensory impressions at all - why can't you just sit there in the dark being conscious? Some would say that consciousness is always self-consciousness - it's only when you are aware of yourself, or perhaps of your own thoughts, that you are really conscious. But it doesn't seem to me that I spend every waking moment thinking about myself, and if every thought I had was sort of labelled 'my thought about x' I'd suspect I needed some kind of therapy for narcissism or paranoia. Some people would say that only explicit thoughts are truly conscious - perhaps only ones expressed in language, but at any rate the sort of thought simple animals can't manage. But if you compare the state of mind of a fish which has just been landed with its state of mind shortly after being clubbed on the head, I reckon there's an awfully strong resemblance to consciousness and unconsciousness.

Of course. But that's the whole point, isn't it? These questions you're raising are the very ones we're concerned with, if we're concerned with consciousness at all.

No, that's my point. If you're a philosopher, maybe these issues are the point. But the rest of us don't really care all that much: what we mainly want to know is how the thing works. Neurologists want to know how the conscious functions of the brain operate, how they go wrong, and how (maybe) they can be put right. AI researchers want to know how to generate or simulate conscious processes, and what interesting results they can get out of it. If they have to get the definitions clear and make relevant distinctions in order to achieve their goals, then well and good - but it isn't the point of what they're doing.

Alright, but they always do have to get the definitions clear, don't they? That has surely been the problem for the more gung-ho school of AI - the people who basically said 'never mind the definitions, pass the soldering iron'. These people thought the philosophical issues were just waffle which could be ignored, but gradually their projects ran into unforeseen problems. Take the 'naive physics' project. The idea there was that you didn't need a deep theory of thought. What you needed to do was write up people's ordinary beliefs about the world in formal guise, throw in a bit of logic (maybe with a few innovations along the way) and you would gradually get the thing to work. You'd develop a system which drew conclusions about the physical world the

same way human beings did, and along the way you would acquire whatever you needed by way of theory. But it didn't happen. The theory never emerged, and it became increasingly clear that the only way to get the programs to work was to 'cheat' by leaving out the difficult bits or building in some of the experimenter's own human common sense about a particular restricted domain.

 I don't think that's an adequate characterisation. I grant you the 'naive physics' thing never really flew in the end, and I'll even agree that the lack of a general theory was part of the problem. But not the lack of a philosophical theory.

On the contrary, it seems to me that the philosophical issues have constantly been a distraction from the real task, drawing researchers from their course and on to the rocks of sterile academic debate. Whenever anyone got a promising approach going, the philosophers would say: ah, yes, but that isn't really consciousness because it doesn't include x variety or cover such and such a form of consciousness. The researchers would get drawn into trying to explain qualia or self-awareness or some other nebulous thing, and their theory would capsize. So what are we going to do about it? Well, my suggestion is that we steal a technique from philosophical argument. This is where you give your opponent a word or the whole field. So if you were arguing about free will, for example, and someone kept saying, yes that's all very well but it doesn't explain what I and most other people mean by free will, you just say: OK, I'll give you the term 'free will' - I'll just accept that if it's defined the way you want, it remains an insoluble mystery. But now let's talk about my concept of, say, 'ffree will', which I'll define the way I want. You can't object that it doesn't match the normal definition, but I'll be able to draw all the interesting conclusions I wanted to draw anyway - just for the price of an extra 'f'.

 So we're going to talk about fconsciousness, is that it? Or cconsciousness?

 No, I thought I'd go for 'konsciousness'. So what I propose is that the scientists say to the philosophers - OK, you have consciousness - do what you like with it. You can go on generating confusions and mysteries about it for as long as you like. Meanwhile, we're going to address our own concept of konsciousness. You can even have an interesting new philosophical debate: is konsciousness consciousness? But pardon us if we don't join you in discussing it.

 And konsciousness is...? Or are you going to leave that on one side while you get on with the programming?

 No - I propose that konsciousness is the name of any process which allows a machine or an organism to produce novel but cogent responses to input stimuli, by using an internal model of the external environment.

 You're helping yourself to rather a lot of stuff there, aren't you? How are you defining 'cogent', for a start?

 There you go again. I don't have to define it. I can recognise it, and that's all I need for my purposes.

 Hmm. Well, I would press you on that, but I see the word 'kogent' looming up, and I think the English language has suffered enough for one day...

Multiple Drafts

1 December 2004

When two hostile traditions independently arrive at the same solution to a given problem, it must be worth a look. In the current JCS, Dieter Teichert suggests that something like this may have happened with the problem of the self: it seems that both Anglo-Saxon analytic philosophy (exemplified here by Daniel Dennett) and its Continental European counterpart (represented by Paul Ricoeur) have come to the conclusion that selves are constructed from narratives; that in the end, we are the stories we tell ourselves about ourselves.

Is the gap between the two traditions really so bad? As a philosophy undergraduate, I seem to remember affecting an airy lack of interest in continental philosophy, and I was not alone. It was alright to have read a bit of Sartre, but I remember the professor once remarking, without regret, "I'm afraid we don't have anyone here who knows about Husserl". In those days I took it for granted that French philosophy would eventually be recognised as merely a branch of literature, while its Anglo-American counterpart had rigorous intellectual standards at least as demanding as those of science. The metaphorical Channel separating the two traditions seemed unbridgeable.

Teichert actually doesn't claim that Dennett and Ricoeur have identical views, merely that there is no conflict. There clearly are some significant differences in context and aspiration. Dennett is interested in causality and clearing up the metaphysics of the self; Ricoeur is primarily concerned with the self as a social and ethical concept. Both, however, have concluded that the self is basically a narrative entity, and that any attempt to give it a free-floating independent status is misguided. Both are reacting against a Cartesian vision of the autonomous, directing self, yet both would, in slightly different ways, stop short of bald scepticism – the self may not be what we thought it was, but it remains in some lesser sense real and well worthy of study.

I think the general point is well made by Teichert, but inevitably one eventually asks – is this interestingly shared view actually correct?

The first point against the narrative self is perhaps its counter-intuitive nature. I certainly seem to myself to have a reality which fictional characters lack. I don't think that if I stopped believing in myself, I should suddenly disappear; nor indeed is it evident to me that I am endlessly narrating myself – or anything. Above all, the idea that my biography is generating me, rather than the other way round, seems distinctly odd. It leaves us with the problem of where, in that case, the biography came from.

The easiest answer is to think of a way in which my story can still, in some sense, be generated by me while at the same time it reciprocally sustains my existence. There is an obvious risk of problematic circularity here.

Ricoeur, it seems, is unbothered by circularity, and actually embraces it. This is relatively easy for him to do, because when one is concerned primarily with social and ethical entities, it isn't necessary to worry too much about their origins: if six people adopt a particular social practice, then the mere fact of their doing so makes it a reality. The fact that the process may involve

some circularity doesn't really matter. Dennett is in a more exposed position, since his metaphysical concerns and aspiration to scientific rigour mean he needs everything to have an appropriate cause, and cannot easily get away with things that pull themselves into existence by their bootlaces. His proposal is that we begin by narratising about others and develop a sense of self when we turn that habit back on ourselves. Recognition of other selves precedes the recognition of our own. This is a little strange, and may still seem intuitively unappealing, but it is not logically problematic.

The argument which (on the whole) convinces me, however, is best seen by considering the intuitive appeal of the narrative self idea. As I've said, the idea seems unlikely in some respects, but in others it undoubtedly does appeal. One reason, I think, is that unlike some philosophical theories, it can be stated in fairly clear everyday terms. Stories are far from being unfamiliar metaphysical abstractions, and any explanation in terms of narratives therefore seems a better explanation than one which uses obscure ideas which themselves need some elucidation. The second reason is that with theories along these lines we really seem to be getting somewhere for once – they do seem to capture some of the mystery of personhood.

The snag is that they may do so in an illegitimate manner. One of the key issues about conscious thought is intentionality – how words or thoughts can be *about* things. Stories, of course, are full of intentionality, and by allowing them into our explanation there is a risk that we are simply smuggling intentionality into our theory without actually explaining it. I don't think either Dennett or Ricoeur is directly guilty of this kind of sleight-of-hand – both have substantial things to say, either about intentionality or the nature of narratives which (if accepted) fill the potential gap. Nevertheless, I do think the tacit presence of intentionality subliminally contributes much of the intuitive appeal to the way they invoke narrative. Without that added appeal, the idea might look very much like a mere category mistake.

Minsky

15 December 2004

Is it time to rehabilitate Marvin Minsky? As a matter of fact, I don't think he was ever dishabilitated (so to speak) but it does seem to be generally felt that there are a couple of black marks against his name. The most widely mentioned count against him and his views is a charge of flagrant over-optimism about the prospects for artificial intelligence. A story which gets quoted over and over again has it that he was so confident about duplicating human cognitive faculties, even way back in the 1970s when the available computing power was still relatively modest, that he gave the task of producing a working vision system to one of his graduate students as a project to sort out over the summer.The story is apocryphal, but one can see why it gets repeated so much. The AI sceptics like it for obvious reasons, and the believers use it to say "I may seem like a gung-ho over-the-top enthusiast, but really my views are quite middle-of-the-road, compared to some people. Look at Marvin Minsky, for example, who once..."Still, there is no doubt that Minsky did at one time predict much more rapid progress than has, in the event materialized. In 1977 he declared that the problem of creating artificial intelligence would be substantially solved within a generation.

The other charge against him is that in 1969, together with Seymour Papert, he gave an unduly negative evaluation of Frank Rosenblatt's 'Perceptron' (an early kind of neural network device which was able to tackle simple tasks such as shape recognition). Their condemnation, based on the single-layer version of the perceptron rather than more complex models, is considered to have led to the effective collapse of Rosenblatt's research project and a long period of eclipse before a new wave of connectionist research came along and claimed him as an unfairly neglected forerunner.

There's something in both these charges, but surely neither is all that damaging? Optimism can be a virtue, without which many long and difficult enterprises could not get started, and Minsky's was really no more starry-eyed than many others. The suggestion of AI within a generation does no more at most than echo Turing's earlier forecast of human-style performance by the end of the century, and although it didn't come true, you would have to be a dark pessimist to deny that there were (and perhaps still are) some encouraging signs.It seems to be true that Minsky and Papert, by focusing on the single-layer perceptron alone, did give an unduly negative view of Rosenblatt's ideas – but if researchers were jailed for giving dismissive accounts of their rivals' work, there wouldn't be many on the loose today. The question we have to ask is why so much notice was apparently taken of Minsky and Papert when a judicious audience would surely have balanced their claims against those of Rosenblatt and his supporters.

I suspect that both Minsky's optimism and his attack on the perceptron should properly be seen as crystallizing in a particularly articulate and trenchant form views which were actually widespread at the time: Minsky was not so much a lonely but influential voice as the most conspicuous and effective exponent of the emergent consensus.

What then, about his own work? I take the most complete expression of his views to be "The Society of Mind". This is an unusual book in several ways – for one thing it is formatted like no

other book I have ever seen, with each page having its own unique heading. It has an upbeat tone, compared to many books in the consciousness field, which tend to be written as much against a particular point of view as for one. It is full of thought-provoking points, and is hard to read quickly or to summarise, not because it is badly or obscurely written (quite the contrary) but because it inspires interesting trains of thought which take time to mull over adequately.

The basic idea is that each simple task is controlled by an agent, a kind of sub-routine. A set of agents which happen to be active when a good result is achieved get linked together by a k-line. The multi-level hierarchical structures which gradually get built up allow complex and highly conditional forms of behaviour to be displayed. Building with blocks is used as an example, starting with simple actions such as picking up a block, and gradually ascending to the point where we debate whether to go on playing a complex block-building game or go off to have lunch instead. Ultimately all mental activity is conducted by structures of this kind.

This is recognizably the way well-designed computer programs work, and it also bears a plausible resemblance to the way we do many things without thinking (when we move our arm we don't think about muscle groups, but somehow somewhere they do get dealt with); but it isn't a very intuitively appealing general model of human thought from an inside point of view. It naturally raises some difficulties about how to ensure that appropriate patterns of behaviour can be developed in novel circumstances. There are many problems which arise if we just leave the agents to slug it out amongst themselves - and large parts of the book are taken up with the interesting solutions Minsky has to offer. The real problem (as always) arises when we want to move out of the toy block world and deal with the howling gale of complexity presented by the real world.

Minsky's solution is frames (often compared with Roger Schank's similar strategy of scripts). We deal with reality through common sense, and common sense is, in essence, a series of sets of default assumptions about given events and circumstances. When we go to a birthday party, we have expectations about presents, guests, hosts, cakes and so on which give us a repertoire of appropriate to deploy and a context in which to respond to unfolding events. Alas, we know that common sense has so far proved harder to systematize than expected – so much so that these days the word 'frame' in a paper on consciousness is highly likely to be part of the dread phrase 'frame problem'.

The idea that mental activity is constituted by a society of agents who themselves are not especially intelligent is an influential one, and Minsky's version of it is well-developed and characteristically trenchant. He has no truck at all with the idea of a central self, which in his eyes is pretty much the same thing as an homunculus, a little man inside your head. Free will, for him, is a delusion which we are unfortunately stuck with. This sceptical attitude certainly cuts out a lot of difficulties, though the net result is perhaps that the theory deals better with unconscious processes than conscious ones. I think the path set out by Minsky stops short of a real solution to the problem of consciousness and probably cannot be extended without some unimaginable new development. That doesn't mean it isn't a worthwhile exercise to stroll along it, however. I don't know whether Minsky himself still hopes for a big breakthrough. On his own website there are a number of interesting papers and some chapters of a new work in progress – on the emotions. A turning aside or a possible new route forward?

Two for One

30 December 2004

Over Christmas, I was looking at an interesting old book - *'The Mechanic's Pocket Dictionary'*, by William Grier. This was the eighth edition, published in Glasgow in 1846. A surprising amount of the book is still an interesting read; the really striking thing is what it reveals about the scientific outlook of the average engineer as late as 1846.

For one thing, Grier starts with an introduction which carefully explains the meanings of the common mathematical symbols '=', '+', '-' and so on. *'Many who are not acquainted with the use of these signs'* he explains, *'are led to believe, that they are employed merely as a parade of mathematical learning, which only serves to make the subject more intricate; but this is a mistake...'* It seems extraordinary now that the average engineer of 1846, someone interested in constucting marine steam engines and other high-tech Victorian equipment, might have regarded basic arithmetical symbols as a pretentious affectation.

Equally surprising is the fact that, according to Grier, some at the time believed that evaporation was due to a chemical reaction between water and air - though Grier himself dismisses the idea. How could mid-nineteenth century engineers have been so confused about the nature of evaporation? Well, one reason may be that, again according to Grier, they had not properly grasped the distinction between heat and temperature, or the precise nature of either.

'It is matter of dispute among philosophers whether the cause of the phenomena of heat be a subtile fluid, capable of penetrating all substances, or a peculiar vibration, rotation, or other kind of motion' he says. It is remarkable that the theory of caloric fluid was apparently still a respectable contender at the time, and scarcely less surprising that engineers then regarded the nature of heat as a matter best left to the philosophers. Grier goes on to describe with great clarity the experiments which show that heat can be reflected and focused by mirrors, a phenomenon, he says, known as 'radiation', and it is hard to believe he had much time for caloric fluid himself.

The case of heat and temperature provides an interesting example of a scientific problem whose chief difficulty lay in drawing a conceptual distinction between two related phenomena which had previously been regarded as the same thing. Personally I can sympathise more readily with the mid-Victorian engineers when I remember my own continuing confusion over the nature of light, where we seem to have a comparable but opposite difficulty - waves or particles? Well, sort of both at the same time, really, whatever that means...

It has, of course, been argued that many of our problems in grasping the nature of consciousness stem from our tendency to treat it as a single phenomenon, when in fact the word refers to two or more quite different things. Perhaps the clearest and most outspoken advocate of this point of view is Ned Block, who refers to consciousness as a 'mongrel' concept, and draws a distinction between *access consciousness* and *phenomenal consciousness*.

Block's way of dividing up consciousness is useful because it captures clearly the distinction which people frequently make; between subjective, qualia-laden experience on the one hand, and relatively comprehensible, computational decision-making processes. This is pretty much the same distinction as the one drawn by Chalmers between the 'hard' and 'easy' problems, for

example. Most people, I think, would accept that it is a distinction worth making, though some would add other divisions, or incorporate Block's point in a more complex hierarchical structure.

But in another way, the distinction is not helpful at all. When we draw a distinction like this, we want it to dispel at least some of the mystery around the topic. Block's approach seems, instead, to leave phenomenal consciousness as mysterious as ever. This is OK if, like Dennett , you want to go on to denounce qualia in all their forms as delusions, but not so good if you want to find ways of naturalising them, as I think most of us do. Moreover, even the 'easy' side of the split is left more mysterious than first appears. Our problem-solving, decision making thoughts may be computational, but they certainly need to be about things, and the mystery of aboutness, or intentionality, is arguably as intractable as that of qualia. It gets slightly less attention mainly because it is, as it were, better camouflaged. All our thoughts are so saturated with intentionality we find it easy to overlook.

A completely different tactic here is to try to make one mystery explain the other. Thus Searle , if I have understood properly, would like to see the roots of intentionality in qualia: the qualia of hungriness is inherently *about* food, and it is from such inherent aboutness that all the other meanings and aboutnesses derive. If this is broadly the right line to be pursuing, then Block's distinction pulls apart the very things we need to bring together.

I'm not absolutely sure that the Searlian strategy is the right one. The general idea of trying to redefine the issues along completely new lines seems appealing however. From some angles the problem of consciousness looks daunting, but it sometimes seems that if only we could realise that we are confusing *x* with the completely separate matter of *y*, or that we need to consider *n* as really just another facet of *p,* the whole thing might suddenly make a new kind of sense. Perhaps we do need a cognitive Einstein who can point out that the mental equivalents of space and time are not really separable after all, but different facets of the underlying reality, interval, or provide some other equally fundamental change of perspective. Then, in 150 years or so, people will find it hard to believe we were still confused about all this.

Theodicy

15 January 2005

There has been a remarkable outbreak of discussion lately on what the recent tsunami implies about God. The pieces I've read have generally started from the observation that such disasters show that God is either not omnipotent or not benevolent; better to be an atheist, they conclude, than accept an enfeebled or malicious Deity. There was even some suggestion, later denied, that the Archbishop of Canterbury was tempted by this line of thought. Now the idea that God is powerless, indifferent, or malevolent is indeed rather scary: but that is obviously a logically feeble reason for rejecting it. Moreover, God's existence is to a large extent beside the point here. When we ask "What kind of God could let this happen?" we're not looking for ontological clarification. Rather the issue is how the fact of sudden arbitrary death and suffering can be accommodated within any tolerable view of human existence in the world; and this needs an answer of some kind even from atheists. God, I think, comes into it mainly because He provides a convenient metaphor with which to address such issues, even if we don't believe in his literal existence. And after all, if Richard Dawkins can talk metaphorically about the selfishness of genes, why should we hold back from talking metaphorically about the coldness of God?

All right then - what kind of God could let this happen? The idea of a spiteful or vicious God is just as difficult to sustain here as the idea of a loving one. An omnipotent God, or even a merely powerful one, could surely do much more to make our lives miserable; and surely He would want to taunt us with open threats and menaces. The real point about God, in fact, is that he does not seem to intervene in the world at all; even 'acts of God' such as earthquakes and tidal waves have clear physical causes and don't seem to require divine causation. We seem to be dealing with the entity a friend of mine used to call *Larvatus* - God in a mask - and the question must be, why does he hide his face from us?

It could be that God is simply mad, or a clown, or on the contrary so far above our understanding that his behaviour remains forever inscrutable. Most likely of all, he could simply be supremely indifferent to us. But the very consistency of his non-intervention tells against these hypotheses. A mad God couldn't observe the restraint necessary to stay out of our way, and an incomprehensible one would surely find that his mysterious movements intersected detectably with human history at regular intervals by sheer chance. The balance of probability must be that the absence is deliberate, that it has a purpose. There is a *project* behind the mask.

At this point, we have to ask - is this enquiry wise? If God wants his project kept secret, had we not better avert our eyes? It seems possible that all discussion of God and his will is disastrously ill-advised, and that religious observance is actually annoying and frustrating to Him. Perhaps Pascal's suggestion that we have nothing to lose by belief in God, and everything to gain, is a tragic inversion of the truth. But without knowing something about God's secret project, it is impossible to be sure. The project might be malign, and the there might be discoverable ways to mitigate the danger. Or God might want us to guess His riddle.

The only clue we have to the nature of the project is this same absence of its sponsor. What project could there be which requires its initiator to exercise no influence on the product? Well - and I expect, gentle reader, you can see this coming - according to some schools of thought that is among the challenges which face anyone who wants to generate artificial consciousness. If the beings we create are to own their own thoughts and have real moral responsibility, they cannot, on this view, be programmed to behave in a given way, no matter how complex the program: they have to arise, presumably through some broadly evolutionary process. This idea resembles a classic explanation of the existence of evil in the world - God wanted us to have free will, so he could not intervene directly either to make us automatically virtuous in the first place or to save us from ourselves when we succumb to temptation or stupidity. The argument is fine as far as it goes, but the snag is that it doesn't seem to go quite far enough. It explains why God might have allowed us to develop as the imperfect beings we are, but it doesn't offer any convincing reason why he could not give us some clear and direct guidance or save us from arbitrary catastrophes. Having undeniable signs of his existence might cramp our style a bit, but it would hardly deprive us of free will or turn us into robots.

Well no, but if we look at things with an existentialist eye, there is a sense in which we constitute ourselves through our acts and the roles we choose. Until we create ourselves by deliberate choices, on this view, we don't really exist on some vital moral plane. C.S.Lewis, J.R.R.Tolkien and their friends seem to have believed that we are called on to be co-creators of the world: perhaps the truth is that we are called on chiefly to be creators of ourselves, and that that process could never be properly carried on in the evident presence of an omniscient and all-controlling God. Perhaps, in other words, God wants us to stop worrying about him and get on with the task of finding better and more substantial selves to be. There can't, however, be any certainty unless and until, unimaginably, the mask comes off.

There can't, however, be any certainty unless and until, unimaginably, the mask comes off. So much by way of metaphor - but is there anything behind the mask after all? It seems to me that the progress of modern cognitive science is actually producing a new and more formidable reason to doubt the existence of God. Generally, atheists have had to rely on the lack of evidence for God's existence, the absence of any particular need for that hypothesis. But as we reach a better understanding of how the property of consciousness is rooted in physical brains (and how it arises from the process of evolution) it becomes harder to see how a non-physical non-evolving being could possibly have it. I don't think the task of accounting for God's consciousness is demonstrably impossible, but it does seem that theists have much more explaining to do than was once the case.

Perhaps, in view of the above, this is a good thing - perhaps, even if God does exist, we are lucky to have such good reasons for atheism.

The CEMI Theory

30 January 2005

 Susan Pockett isn't the only person who thinks consciousness is basically an electromagnetic field. Johnjoe McFadden put forward a similar theory back in 2000 and his version has some advantages. There is a short account of the theory on his own website at the University of Surrey, with the two papers which flesh it out more fully. In his view, the endogenous electromagnetic (em) field produced by the human brain contains the same information as the neurons, but in a form which is analogue, integrated and distributed; all characteristics which seem to be shared by consciousness and not by the digital, discrete activity of the neurons themselves. His version of the theory steers well clear of epiphenomenalism and all its problems: the em field arises from neuronal activity in a normal way and has straightforward causal effects on the firing of neurons in its turn.

He concedes that the very small charges involved may mean that quantum effects are relevant; but quantum mechanics plays no special part in the theory and he does not rely on strange quantum effects to explain the strange properties of consciousness. It does seem, however, that his theory can clarify a lot of different problems. The distinction between conscious and unconscious action, for example, is simply a matter of whether the em field was, or was not, playing a decisive role at the time. As we drive along the road, our over-learned responses produce robust patterns of firing in the neurons which require no intervention from the em field; but if we hit a novel situation our neuronal response becomes confused and less co-ordinated, and the subtle influence of the em field takes over - we begin to think about what we are doing again.¶

 That's all very well - but we don't stop thinking altogether while we're driving. If our mind wanders, we start thinking about something else. If McFadden is right, these other thoughts must be coming from neuronal activity too, mustn't they? So he thinks that in that case the electromagnetic field will be most influenced, not by the strongest patterns of firing, but by weaker ones? That doesn't make much sense. If his theory were right, it wouldn't be that our mind wanders - we'd suddenly find that our actions had gone terrifyingly out of conscious control every time a habitual pattern of behaviour kicked in.

 No, no. How could we be terrified by something which, by definition, we have stopped paying attention to? Besides, the contents of consciousness are not decided merely by the largest electrical influence - I'll explain that in a moment. The theory also offers a convincing explanation of qualia. It would be perfectly possible for our brains to run unconsciously through pure neuronal computation, taking account of sensory inputs and modifying behaviour accordingly - but it is the extra buzz of em activity which gives them their phenomenal qualities, engaging them in the process of consciousness. This view has the advantage of setting qualia apart from the routine causality of mental processes without rendering them irrelevant.

McFadden thinks fading and absent qualia are perfectly possible; but at the same time a person or robot without qualia would be readily distinguishable from a fully conscious person. McFadden doesn't think, incidentally, that artificial consciousness is impossible, if the machine were constructed in the right way. There are fascinating results here from Sussex University. A neural network was trained to distinguish two tones: once trained it emerged that some of the cells which were essential to performing the task were not actually connected to the rest! The only explanation is that they were contributing to the performance of the network through some field effect - very much as the em field hypothetically would do. The implication is that this network had a dim, restricted form of phenomenal experience.


I simply don't see why we should assume that an electromagnetic 'buzz' has anything more to do with qualia and actual experiences than the firing of neurons (or any other physical process). The problem is that physical processes and real experiences are as different as chalk and cheese, and substituting one physical process for another makes no difference whatever.


If you take that line, you're simply making the problem unanswerable by definition. But anyway, McFadden's theory also offers an obvious solution to the issue of free will. The problem with free actions is that they seem to come out of nowhere; well, the truth is that come out of the em field. They really are exceptions to the underlying neuronal causal process, but there's nothing spooky or magic about them: they are perfectly normal results of normal physical processes.¶"Blandula"The theory seems incomplete to me. After all, the brain (and the rest of the body) is full of activity which generates electromagnetic fields. Why should only some of them, on this reading, be conscious? For that matter, why aren't ambient electromagnetic fields, like that of the Earth itself, conscious? Why don't television broadcasts have minds of their own, and given the strength of them, why don't they over-ride our conscious thoughts? At least Susan Pockett acknowledges that only certain kinds of patterns of activity can be conscious.


First, it's obvious that a degree of complexity is necessary: no one supposes a simple electromagnetic field is automatically conscious. Second, the brain is actually rather well insulated from outside fields. Where magnetic fields are strong enough and targeted in the right place, they undoubtedly do disrupt mental activity. Although there is a difference here between Pockett and McFadden, he accepts that not all field effects arising from the brain are the same. He specifies that only those which eventually influence motor neurons are to be regarded as conscious.


Why only those? it seems a very arbitrary distinction. And at the time some field event takes place, you can't tell for sure whether it will set off a chain of events that impinges on motor neurons, can you? But it must either be conscious or not at the time it happens, surely?


Yes, of course. The distinction is a practical one which could no doubt be sharpened up theoretically: the basic point is clearly that we're talking about reportable processes in some sense. But we still haven't touched on possibly the strongest advantage of the theory, namely that it also deals with the binding problem. It has always been a mystery how the different bits of

data from different senses get bound together into a coherent, consistent account of reality: but if there is an overarching em field which picks up neuronal influences from all the senses and transmits the combined result instantly over the brain, the solution is clear.

 I don't see that. Part of the binding problem is how processes which run at different speeds are co-ordinated: how does an instant field cope with that? And it seems to me that a single field would bind things together too much: if that's where our view of reality comes from it would all be hopelessly melted together in an incoherent buzz. It is of the essence that the right things get bound together in the right way, not just instantly smudged together.

What really puts me off the theory, though, is that it seems like another blow in favour of electricity. We've always suffered from treating the brain as if it were made of copper wire, when it's perfect clear that subtle chemical effects are absolutely crucial. In fact, I'd say one of our big problems here is that we draw a major distinction between physics and chemistry which Nature simply doesn't recognise.

 There's an element of truth in that, but I think McFadden's theory is a corrective to the 'copper-wire' view, not a reinforcement of it. He points out that these effects exist - ephaptic coupling of nerve activity through electromagnetic fields is an established fact - and surely they need to be considered.

 The trouble with those ephaptic effects is that, as I understand it, they are mainly associated with disorders, like tinnitus. They're like the sort of interference you would sometimes get between old-fashioned phone wires because of induction; unwanted signals in a system which was designed to work without them. After all, neurons go to great lengths to connect up with each other, often at great distances; surely direct effects from an em field are just going to be unhelpful noise.

The bottom line here is that McFadden (and Pockett) want electromagnetism to do what the spirit does in traditional accounts, but it just isn't up to the job.

 I see it as a strength of the theory that it sits well with intuitive and traditional ideas about the way the mind works, while requiring nothing but ordinary mainstream science to sustain it. One of its virtues is that it is amenable to experimental testing, and I have no doubt that over time evidence will accumulate. What I'd really like to see is an AI project in parallel, seeing whether there aren't practical advantages to the kind of set-up McFadden describes. Unfortunately I don't think this is happening at the moment.

Free

15 February 2005


That old favourite, free will versus determinism, seems to be in the air again. The other day I was reading James Anderson's 'Perspex' theory (*more on this another time*) and came across his formulation that: *'An agent's will is free if it is not willed by another agent.'* Then the latest issue of the JCS is devoted to David Hodgson's attempt to give a philosophical justification for something like the common sense view of free will.

Anderson's definition, as quoted above, seems to need some revision. Although not being subject to someone else's will is certainly a condition of free action, it isn't the case that someone else can deprive me of freedom just by willing me to do whatever I am actually doing. My friend may fervently hope I have bought a particular book she wants to read; that doesn't mean that, as I stand in the bookshop making that particular purchase, I have no freedom of action. In general, however, from the quick look I've had, Anderson seems to make a fairly sensible compatibilist case.

Hodgson, on the contrary, is out for freedom red in tooth and claw: a freedom incompatible with determinism. I admire his pluck while deploring his judgement. I think it's a quixotic enterprise, but anyone who sets out to defend the common sense point of view has my sympathy.


Is radical libertarianism really the common sense view? Frankly, I'm not sure. The common sense view isn't actually a crisply articulated philosophical thesis, is it? That's why we have to do philosophy in the first place, because the common sense view is unclear in itself.


Well, I suppose I'll grant you that, though I think radical determinism is pretty far from common sense. Hodgson sets out nine theses. Probably the key ones are the first and fifth. The first says that there are situations in which the laws of nature allow two possible states to follow a choice, rather than just one. The fifth says that people can make effective non-random choices between the alternatives, influenced but not determined by any rules or reasons. I feel there's an element of duplication here - we've got an indeterminacy in the world and an indeterminacy in the mind. If you allow me a mental capacity to make effective non-random, non-determined choices, I think I can run you up a perfectly good system of free will with that alone - I don't need an additional indeterminacy in the external world. But perhaps I've got the wrong end of the stick.

I think most people will find it hard to accept that there can, strictly, be two possible sequels to any given situation. Hodgson invokes quantum mechanics, but even if we accept that there's an indeterminacy there, that in no way authorises him to assume macroscopic indeterminacies of the sort he needs.

 I think it's hard to see how there can be any real indeterminacy in the world. There is a deep metaphysical principle that nothing happens unless it must; and that what does happen is the minimum necessary to satisfy the laws of physics. If it were otherwise, anything could happen at any time, and the world would be arbitrary, if not actually null.

 But the laws of physics might be such that they allow two or more different minimum outcomes, mightn't they? Anyway, my main intention was not to discuss Hodgson.

 So what, then - your own half-baked ideas? I can do without another discussion of the supposedly wonderful, unique, magic human property of 'free' action. All you really need to know is that determinism is true. And I repeat, it has to be true. Even if all our current science somehow turned out to be wrong, the Universe has to follow some set of rules - some unambiguous set of rules. Otherwise it would be incoherent, unmotivated, or both. And if it follows rules, it's deterministic.

In fact, here's an even simpler argument. Event A happens. At all previous times, then, it was true that event A was going to happen. So event A was already determined at all previous times.

That is all. Tell the fat lady to start singing.

 Don't jump to conclusions. In fact, I want to argue that the problem is illusory; that this is one of those issues where the real challenge is not finding the answer, but working out why people think there's a problem. Now, the problem is supposed to be a contradiction between two sentences like the following.

- My action A happened because of the preceding physical causes that gave rise to it.
- My action A was freely chosen by me.

In fact, there is no contradiction. Sensible people since, oh, St Augustine at least, have been patiently pointing this out, but it never makes any difference. People always feel, and probably always will, that there is a problem here. I think...

 Whoa! The contradiction is obvious. If sentence 1 is true, then action A had to happen - you could not have done otherwise. If sentence 2 is true, action A did not have to happen, and you could have done otherwise. Contradictions don't come much clearer than that. It's just that sentence 2, strictly interpreted, is false.

 No, no. You see, discussions of what could have happened always depend on implicit but limited assumptions about the circumstances. That's what makes them worthwhile, because they shed light on the particular facets of the circumstances which are assumed to be under consideration. If you tell somebody you could jump a fence, you're telling them about your own state of fitness and training, not making any kind of prediction about whether the course of your life will ever, in fact, lead you to jump the fence. If someone asked you whether you could jump the fence and you said "No, because I don't intend to and therefore won't." they

would think you were misunderstanding (or perhaps rhetorically countering a perceived suggestion). But that's the kind of stance you take up when you go off on your determinist tack. In effect, you just refuse to talk about possibility. You insist on the whole history of the Universe being implicitly specified down to the movement of every atom. Then you say; given the state of the Universe, only one thing could have happened. But that's vacuous. You're just saying 'given that everything was the way it was, then things are going to be the way they're going to be'.

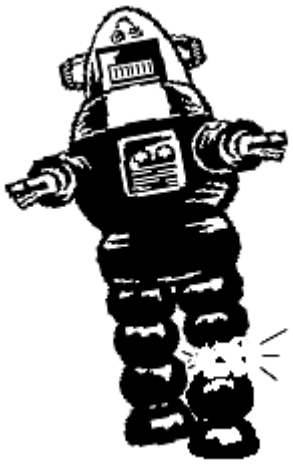

So really, what you're offering us here is yet another version of stale old compatibilism. We're not really free, but don't worry, we can redefine freedom to mean something less demanding?


I'm not offering another version of compatibilism: what I'm offering is, if you like, non-specific compatibilism. This is one of the reasons, I think, that people have difficulty accepting that there isn't a problem. Compatibilists search for a single formulation, a single definition of freedom which will cover all cases - but there is no such universal version. When we talk about being free, we just mean free of the particular constraints which are implicitly under consideration at the moment. We may be free because we're not in jail, because we have the money to do what we want, or because we haven't been brainwashed. Absolute freedom doesn't make sense; only freedom relative to particular projects and circumstances.


I think you're wrong, but this is a healthy development - you've given up one of the supposedly unique features of human consciousness. After all, my computer is free in relation to certain projects and in certain circumstances. It's free to play me a tune if the speakers are plugged in; free to store my files if it has enough space for them. So by your reckoning, I'm entitled to say that it, too, has free will?

Brittle

20 February 2005



A recent paper by Michael L Anderson and Donald R Perlis offers a new solution to the problem of brittleness. Brittleness, in a robot or program, is an inability to cope with unexpected developments, and the paper, no doubt rightly, says it is arguably the most important problem in Artificial Intelligence, perhaps in computer systems generally. Anderson and Perlis quote the example of a contestant in the DARPA Grand Challenge, in which robots had to negotiate their way through a real-world journey. The hapless specimen in question drove into a fence it could not see and then continued attempting to move forward for the rest of the contest. Perhaps better sensors would have helped, but sullenly persisting in a course which isn't getting anywhere is poor behaviour for a supposedly intelligent machine. The proposed solution is the *Metacognitive Loop*; roughly speaking, a mechanism whereby the robot or program can notice whether or not it is achieving its goals. If not, it can then try something else, whether a carefully considered Plan B, or more or less random variations in behaviour, until it gets better results.

Clearly what we are dealing with here is our old friend (or enemy) the frame problem: in fact, although they never use that phrase, Anderson and Perlis cite the paper by McCarthy and Hayes which first clearly identified the problem, and in some ways their work is in harmony with McCarthy's view that the problem would probably be resolved by new (non-monotonic) approaches to logic.

The subject of consciousness notoriously exists at the junction of many different fields: psychology, philosophy, neurology and artificial intelligence to name just some. The frame problem, I think, is a leading example of how different specialists can have, and retain, a very different view of the same problem with surprisingly little cross-fertilisation. Since the issue first came to light, AI specialists have seen it as a persistent, annoying problem, but one capable of being dealt with tactically in the short term and by prudent measures in the longer run. Philosophers, in contrast, have tended to see it as a fundamental issue: one which requires a completely new approach or insights, or even rules out all possibility of artificial consciousness. Even Daniel Dennett, perhaps the philosopher most supportive of AI, has given a classic exposition *('Cognitive Wheels- the Frame Problem of AI')* of his own version of the frame problem without, so far as I know, offering any solution.

It must be acknowledged that on the face of it the AI fraternity frequently seem to have the more reasonable view. Some philosophers present such a gloomy case that we seem almost obliged to disbelieve in human consciousness, never mind the artificial variety; and approaches like that of Anderson and Perlis seem extremely reasonable and pragmatic.

They identify three distinct problems with 'common sense' reasoning. One is 'slippage' - the facts don't hold still over time, but are liable to change while you are reasoning about them, or clinging to outdated conclusions. The second is the *'KR mismatch'* problem in which the representational conventions used impose unhelpful (but unavoidable) limitations on the ability to deal with circumstances which don't match them - especially those which give rise to novel

expressions not catered for in the original scheme. The third issue is how to cope with contradictions. Contradictions in the system's data are inevitable, but making decisions in their presence is obviously problematic, especially for a system which operates by classical rules of inference, since in classical logic there is a valid inference from a contradiction to *anything at all*.

There are two steps in the proposed solution. First, the paper advocates *'active logic'*. In this approach, instead of logical inferences taking place instantly (or in an eternal Platonic realm), they follow a chronological sequence. The conclusion comes a moment after the premises, not at the same moment. This has two benign consequences. The system is not obliged to consider all the implications of every piece of data - surely an infinite task in principle anyway - and it can tolerate contradictory premises because it may not happen to be thinking about the two sides of the contradiction at any given moment. Where there is a real problem, the system will eventually derive a direct contradiction, which then serves as a useful warning that something is going wrong.

This part of the argument seems dubious. In the first place, if contradictions are signs of a problem, allowing them to remain unresolved in the heart of the system is surely dangerous; by the time a direct, explicit contradiction is noticed, the damage may have been done. Second, is it actually safe to assume that a direct contradiction is necessarily a problem? In ordinary thought that may be so, but in formal logic contradictions are a useful if not indispensable part of the process of reasoning: without them many proofs are going to be hard to derive.

The second step is the metacognitive loop, which seems unassailably sensible. The basic architecture is composed of two modules, one training the other, with a third overseeing both. This third module is what does the monitoring of performance against expectation, and steps in with new strategies and retraining where necessary.

The authors propose three potential projects which the metacognitive loop might enable: a robotic St Bernard, able to render basic assistance in unpredictable emergency situations; a Mars explorer, which might be able to cope well with finding things its designers had not expected; and finally a *GePurS* - General Purpose Scout, designed to operate in many different environments, with a large reservoir of experience, much of it redundant in any given situation, but lending a particular flexibility and generality to its cognitive abilities. Indeed, such a creature would surely be intelligent on an entirely new level, comparable with sophisticated animals or even human beings. They don't, of course, predict the imminent achievement of such high-grade performance.

So, is the road ahead clear? It is interesting to compare the example of the misguided DARPA challenge robot with the one described in Dennett's essay. Dennett's robot is set to retrieve its spare battery from a room where a bomb is due to go off, but singularly fails to take account of the known fact that the bomb is on the same trolley as the battery. The DARPA robot fails its task in a clear way - it isn't moving towards its goal - whereas Dennett's unfortunate robot actually succeeds in getting the battery out: failure arrives out of a blue sky from an unforeseen factor - the fact that the bomb has come along for the ride. It's hard not to feel that metacognitive loops might help the DARPA victim but wouldn't do much for Dennett's robot. If you have a module monitoring success, it has to be able to recognise success when it sees it, and this is not a trivial task. Even the DARPA robot may have had, for example, a speedometer which told it its wheels were revolving in fine style, oblivious of the fact that they weren't moving it anywhere. So long as your robot relies on logic, the only way for it to assess success is to perform a number of

tests set up by the designer: but that means the designer has to anticipate everything which might go wrong - the very problem we started with in the first place.

The relative weakness of formal logic as a tool for reasoning about the world is, of course, well known - and known to Anderson and Perlis, whose proposals are designed to extend its efficacy. But the AI practitioners seem always to under-rate the severity of the problem, perhaps because, when all's said and done, formal logic is still the only real tool available. Aaron Sloman, long ago, argued very persuasively that some kind of analogical reasoning is needed to make good the deficits of old-fashioned logic. There are many problems attached to processes based on analogies, and the development of a suitable formalism is itself a daunting task, requiring a breakthrough of considerable proportions. Nevertheless, it's hard not to feel that he was, broadly at least, on to something.

How *your* mind works, maybe

2 March 2005

Carl Zimmer's ever-interesting blog is featuring a two-part discussion of a spat which has arisen between, on the one hand, Noam Chomsky, with Mark Hauser and Tecumseh Fitch, and on the other, Steven Pinker and Ray Jackendoff. It has to do with the evolution of language: in a very bare summary, Chomsky and co. suggest that most of the necessary abilities were already present in pre-human animals, and that possibly only one additional element, a special capacity to deal with recursion, was needed to get human speech up and running. Pinker and Jackendoff insist it is all much more complex than that, and that speech evolved progressively under selective pressure in much the same way as any other adaptation. Chomsky and co retort that they have lost the plot and are out of date. It is remarkable how bad-tempered disagreements about evolution sometimes get, and it appears that feelings are running rather high here.

There are surely plenty of reasons to stay calm in the current case, though. For one thing the issue is potentially amenable to resolution by further empirical evidence - so why shout about it now? It seems debatable, in fact, whether we yet understand either language or human evolution well enough to tackle such issues on any but the most speculative level. Moreover, how much is really at stake? Although the course of human evolution can never be regarded as trivial, this particular dispute does not seem to hold out the likelihood of any major paradigm shifts in our understanding.

Finally, as has been exhaustively established by science, all theories about the origin of language are false, and those who propose them are wasting their time (admittedly, we have wasted a certain amount of time ourselves).

It seems a good enough reason, anyway, to look again at Steven Pinker's theory of consciousness. The problem, of course, is that he notoriously *has* no theory of consciousness. *The Language Instinct*, his first book, is really excellent: readable and entertaining, informative and stimulating all at the same time; full of expert and overdue debunkings and clear expositions of the ideas of, well, Chomsky in the main, as it happens. *How the Mind Works*, his second, was naturally awaited with keen interest, but proved to be one of those books with titles that claim a little too much. Pinker professed himself stumped by the more difficult aspects of consciousness: in his view the brain was a computer and a product of evolution, but that was about as far as he could go. Whether you agree with them or not, these are not exciting new theses. Perhaps the main point of the book, instead, was to make the case which Pinker has championed again on many subsequent occasions - that there is such a thing as human nature: that much of the way we behave is determined by hereditary predispositions with a recognisable evolutionary rationale. Armed with this insight, it seems, much about human culture and society which was previously shrouded in mystery is helpfully illuminated. At times Pinker seemed to lack all faith in the ability of false but fashionable beliefs to die off of their own accord, but an awful lot of what he said on this occasion, and again, more grimly, in *The Blank Slate*, was true and valuable, and well worth reading.

But whereas in the earlier book he had cantered across the landscape like a mountain goat, here he seemed to lose his footing quite badly once or twice. Part of the problem is surely the variable quality of the sociobiology being put forward: while some of it is undoubtedly thought-provoking, some resembles banal platitudes too closely for comfort. When we are told that men are more promiscuous than women, or that sleeping at night is a universal feature of human culture, we can be forgiven a small sigh.

At times I found myself visualising what might happen if the boot were on another foot and a party of guerrilla novelists had broken into a convention of, say, physicists. Seizing the rostrum, they begin explaining the results of their own research.

"While you physicists have been having your whiffly conversations about quarks and string and... stuff," shouts one, "we novelists having been doing hard research. Here's some real physics for you. Heavy things sink; light things float! Metal doesn't burn! Empty things are lighter than full things!"
"Well," says a physicist mildly, "if you had a lead bucket and..."
"Have you done the experiment?" demands the angry novelist, "No? Well, we have - several times!"
"Look," says the physicist, with infinite patience, " I see where you're coming from, but trust me, we've kind of covered this ground, and it's actually not quite as simple as you think..."

In Pinker's case, religion and the arts seem to be particular blind spots. At one point he suggests:

"Most people would lose their taste for a musical recording if they learned it was being sold at supermarket checkout counters..."

Really? By this reasoning it must be universally acknowledged that Etherege is better than Shakespeare, and Buxtehude greater than Bach, simply because their work is more difficult to find. Of course it is true that snobbery, competition for prestige, shows of status and so on influence the arts. But they also influence the course of science. Pinker would never suggest that this *explains* science, however, because he understands science, and knows that however great a factor social competition may be, it isn't what science is ultimately about. And of course it isn't what art or religion are about, either. So, as it turns out, Pinker does have a theory about consciousness - namely that it doesn't matter as much as we previously thought. People build large structures to live in, and so do ants and termites; certainly there are differences, but perhaps we may have over-rated them. This seems an unattractive line of thinking to me. The fact that human culture and patterns of behaviour tend, (like the instinctive behaviour of animals) to be compatible with the survival and reproduction of those who adopt them is surely one of the less surprising and interesting things to be said about them.

But now perhaps I had better take my own advice and stop shouting.

Sony Qualia Man

10 March 2005


I was interested to see that Sony has been taking a close interest in the issue of qualia (the ineffable subjective qualities of experiences, the actual redness of red, for example which inevitably get left out of the scientific account of perception), and in fact has named its top-level range of equipment (launched in the US about a year ago) after it.

I suppose Sony *is* in the qualia business, though I'd never thought of it that way before. It raises the prospect of a gradual popularisation of the term - in a few years, nervous Hollywood agents pitching their projects to cynical moguls will probably be insisting that their film has 'strong qualia' as well as being 'high concept'. When Sony get round to David Chalmers, will they sue him, sponsor him, or option his book? In fact, they have an intellectual in charge of their qualia project already, in the shape of Ken Mogi.

It's rather nice to see that Sony is sponsoring his research as 'conceptor' of a 'qualia movement' within the company. As if one unexpected twist on the idea of qualia weren't enough, Mogi has issued the Qualia Manifesto, which gives the whole thing an aesthetic, almost an ethical turn. The Manifesto is not, as you might expect, a declaration of belief in the reality of qualia so much as a 'fundamentalist ' call for more of them.


Mogi has a kind of epicurean view of the issue: everyone, he suggests, enjoys seeking out new qualia, and (epistemological issues on one side) sharing them with others. We want more, better, newer qualia to be made available to improve all our lives. It's easy to see how this message might appeal to Sony, (though Ken also believes qualia are not easily marketable), but it is a rather unexpected angle, and I must admit it made me think about some fairly basic issues which I had hardly noticed before.

Mogi, for one thing, seems to assume that all qualia are inherently good; the taste of fine French wine, the majestic sight of the Grand Canyon - it's simply a question of how many we can manage to experience. But surely many qualia are unpleasant - the qualia of clashing colours or stomach-turning smells?


Up until now, I have had a vague half-conscious assumption that there must, in fact, be as many bad as good qualia, all lying on some ineffable linear scale of pleasantness. Now I come to think about it, however, I see no particular reason why the bads and the goods should balance. It could be that good subjective experiences occupy a narrow band in the middle of the imaginary scale, with the infinite realms of too much and too little stretching away on either side. The seemingly uncategorisable variation of qualia actually suggests that there may in fact be no orderly scale at all, and no way of assessing whether the overall cosmic phenomenal balance tilts towards pleasure or pain.

In fact, based on introspective examination of a few qualia, I have come to the conclusion that none of the foregoing alternatives is correct: in fact, qualia in themselves must, I think, be neutral, neither positive nor negative. It is only the pain or pleasure they provoke which makes

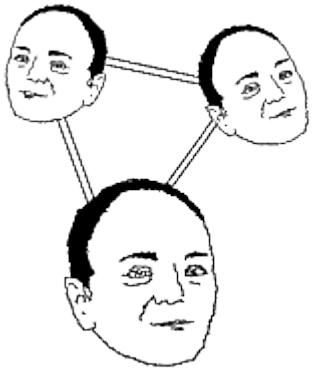
them seem bad or good. Pain itself is generally regarded as a quale or a family of qualia, and I suppose the same could reasonably be said of pleasure; but they are both conative in a way that other qualia are not: they make you do, or want to do things (get your finger out of the electricity socket; stuff the cream cake in your mouth) - in fact, the conation may very well be all there is to them. Pure pain, on that view, is pure aversion, though we rarely if ever experience it in the absence of fear, anger, awareness of some specific bodily damage, and so on. I do remember once having mains electricity routed through my finger as the result of my own stupidity, and apart from the mild buzzing caused by the alternation of the current, the only phenomenological content of the experience was a very keen desire for it to stop as soon as possible.


So is Mogi right? Setting aside the basic desire to avoid pain and seek out pleasure, should we want more qualia? If the majority of them are neutral in themselves, why should we? I can understand that people might want to have more pleasurable experiences - seeing new sites of natural beauty, for example - but would we want to go on seeing ever more of an endless row of identical concrete bunkers? That's not quite a fair comparison, however, because Ken stresses the value of new, never-before-experienced qualia. It might be that the enlargement of our experience does tend to enlarge us, or our minds, in a way which is desirable quite apart from the direct pleasure involved. Perhaps Ken *is* on to something interesting.


Since qualia do not exist, that seems unlikely. *Nothing* need get left out of the scientific account of perception. Wish I could afford some of that hardware, though.

Deep Thoughts

20 March 2005



Gregg Rosenberg's book "A Place for Consciousness - probing the deep structure of the natural world " is the most ambitious metaphysical project I have come across for a long time. Not only does it offer a new view of consciousness, it suggests a new and elaborate theory of causality, and a new kind of ontology to go with it. It pushes metaphysical speculation boldly into regions which science has pretty much regarded as its own for many years, and brusquely denies the completeness of the physical account. It also embraces philosophical positions which most would regard as untenable: in this, it recalls David Chalmers, and indeed Rosenberg positions himself as operating in a kind of post-Chalmers context with some Chalmerian (Chelmersian?) leanings.

A breath of fresh air, then? I think so, though I couldn't sign up unreservedly to the theory, and I have particular reservations about the elaborate apparatus which Rosenberg proposes to deal with causality.

Anyway, what's it all about? The basis of the theory is the view that conscious experience - qualia in particular - cannot be satisfactorily accommodated within the physicalist account. Rosenberg proposes an analogy with Conway's "Game of Life". In this game, we have a world consisting of an indefinitely large grid. Each cell can be "off" or "on". Some simple rules about adjacent cells determine, for each successive state of the world, which cells will be on or off. It turns out that this simple set-up gives rise to patterns which evolve and behave in complex and interesting ways. We can even construct a huge pattern which acts as a Turing machine.

Now, says Rosenberg, in the Life world, there is nothing but bare differences. You may be able to generate hugely complex entities within the Life world - perhaps even life itself: but there's no way these bare differences could entail subjective experience. Yet subjective experience is undeniable - qualia are an observable fact. Now when you get right down to it, the world sketched out by physics is also a matter of bare differences. The fundamentals are a little more complex than in the Life world, but in the end you come down to a similar kind of contentless data. It follows that conscious experience is not entailed by physics, and it must therefore be entailed by something else.

The Game of Life is fascinating and instructive, but I think there's a bit of trickery going on here. Rosenberg only ever talks about small sections of the Life world - of course you can't do much with a grid the size of a chess board. He mentions the Turing machine, which requires a huge grid, of course, but even that is a minute fraction of the size of the array you'd need to model the full working detail of even one human brain. The Life grid needed for even a small community of people just beggars the imagination, and at that point it ceases to be convincing that the thing is too simple to support the kind of mental experience we normally have.

At the end of the day, it's just another appeal to intuition we're dealing with here. Rosenberg insists that he's giving real evidence, but if all you have for evidence is the way something looks to you, I say we're just sharing intuitions, and mine are different from his.

 I don't like the argument much, either, but for different reasons. I see why you're inclined to reject subjective evidence - "subjective" has always been a derogatory term in science. But if it's the nature of phenomenal experience we're dealing with, how can you disallow subjective evidence?

Still, I don't think it's legitimate to argue that what's true of a simple game must be true of reality, too. Life is a world only metaphorically; it seems doubtful to me that such a world is really possible (even in the required, fantastically outré, sense of the word "possible"). At least, if it's to be real, there are going to be some problems about preserving identities between the successive, independent moments which constitute time in the game. And that's one of the key points: the Life world is, by specification, a discrete-state world consisting of a binary grid. It is completely computational. The real world, by contrast, is messily continuous and full of non-computable stuff. This is particularly relevant because (according to me and many others) consciousness and qualia are among those non-computable features. Arguing from Life to reality just begs the question.

However, as a matter of fact, I think Rosenberg's incredulity is justified. How can mere physical facts entail subjective experience?

 It's obvious that physical facts entail mental facts. Rosenberg chooses to defend his views with some very subtle and sophisticated argumentation about esoteric philosophical points, but we're not dealing with esoteric matters here. A smack round the head with a spade is a physical fact, and entails plenty of consequences for your so-called "qualia". Less brutally, having a red object stuck in front of my eyes in good lighting conditions entails an experience of redness in my mind. I know there's plenty of scope for quibbling about the exact nature of the counterfactual conditionals and all that, but at the end of the day are you going to say there's no entailment between the physical facts and mental experiences? That's going to disconnect your experiences from the world altogether, isn't it? You could take shelter in a restricted sense of "entailment", but to his credit, Rosenberg quite rightly argues that we need to understand entailment in a broad sense here - certainly much wider than formal or logical entailment, anyway.

But look - the last century or so has seen an accelerating accumulation of experimental results which show just how closely our mental life depends on the physical operation of our brain: yet none of this seems to have impinged on Rosenberg. He argues from a perspective that is almost medieval. We know better than that.

 No, no! You talk as though he were arguing for dualism. Rosenberg isn't trying to disconnect the physical and mental worlds: he's just pointing out that there is more to be said about the world than the bare facts of physics. Remember this is pure physics we're talking about. I think any unprejudiced observer, even a thorough materialist, would accept that there are aspects of the world which cannot be captured by talking about basic quantum physics. I realise some of the arguments are a bit sophisticated for your taste, but isn't it true, as

Rosenberg says, that if conscious experience were directly entailed by simple physics, it would be, in some sense, a free lunch? The idea of getting an additional phenomenal world for free in this sense really is hard to swallow, I think, and I find the argument on this persuasive.

Anyway, the second stage of the argument is a consideration of panexperientialism. This is, if you like, a weaker version of panpsychism. It's not that everything has a mind of its own, but rather that everything has a limited share of simple experience. Rosenberg distinguishes between the kind of full subjective experience human beings have, complex and infused with cognition, and the kind of tiny spark of subjectivity an inanimate particle might be thought to have. Ultimately, his theory sets out a way of building higher-level entities out of these tiny sparks.

There's a curious use here of Ned Block's Chinese Nation argument. Block proposed that each Chinese citizen could be set to reproducing the behaviour of single neuron in a brain: would the resulting higher-level entity really be conscious? (It has been pointed out that in fact the population of China is nothing like as large as the number of neurons in a human brain, and the example has an unfortunate racial tinge, especially taken in conjunction with Searle's Chinese room argument.) Rosenberg says yes, and uses the argument to show that experiencing entities could exist on several levels. For reasons which are not completely clear to me, he regards this as a problem - if experience can happen at different levels, he feels the only logical stopping points are either consciousness as a property of every atom, or consciousness as a property only of the cosmos as a whole. He suggest that there is a "boundary problem" about why consciousness has the limits it does. The existence of consciousness at a middle level therefore needs explanation - but surely the argument actually shows that it could equally well exist at various levels? Rosenberg, of course, ultimately wants to offer an explanation of how low-level experience can be built up into higher-level structures.


Poor old Occam must be spinning in his grave at this point: but can I just spool the argument back for a moment? Rosenberg began by arguing that nothing like bare differences in the world could entail the fantasmagoria of subjective experience, right? But now - and only now - he starts to suggest that subjective experience might be made up of minute glints of experience. Now it seems to me that these tiny sparks of experience look a lot more like the kind of thing you might get from bare differences, and if Rosenberg had started with them his argument would have looked much less persuasive.

There's something a bit shifty about his discussion of panexperientialism, anyway. I'd like it a lot better if he came out thumping the table and declaring that panexperientialism is true, and here's why. Instead, he sort of argues that there's no reason why it couldn't be true - maybe it's even likely? The real reason he wants us to accept the plausibility of panexperientialism is that he wants it as the foundation for his theory, but in itself there are lots of reasons to dismiss it. One, of course, is that it adds an enormous number of experiencing entities to the world for no particular reason, and hence offends against parsimony - though parsimony doesn't seem to be a principle Rosenberg values very much. Second, once again, we know quite well that the functional properties of the brain are closely associated with our ability to have experiences - even quite simple ones. A certain minimum of structure is necessary to have those functional qualities, and single particles certainly don't have that minimum. Rosenberg argues against functionalism itself, but you don't have to think that functional properties constitute consciousness in order to see that it depends on them.

 I think it's true that Rosenberg basically wants panexperientialism for the sake of the theory he builds on it, but why not? If the theory accounts for consciousness and causality, then it's well worth it.

The next step in the argument, in any case, is an assault on causality. Rosenberg launches an attack on what he characterises as Humean views. I have to say that Hume comes over here as a dogmatic figure rather at odds with the gently devastating agnostic I'm familiar with, but causality undoubtedly remains one of the great mysteries. Rosenberg wants us to unscramble some of our assumptions and stop thinking purely in terms of causal responsibility. He proposes the more general idea of causal significance and wants to deal separately with effective and receptive causal properties. The physical account, he suggests, deals only with effective properties, and is hence one-sided.

 Yes - according to him, we're all talking about effective causal properties. Actually, I think that's another mediaevalism, and most of us are not talking about causal properties at all - they sound worryingly vitalistic to me. In another place Rosenberg points out, more accurately, I think, that the account given by physics doesn't actually make use of the concepts of cause and effect as such - it simply describes certain regularities in the space-time continuum. Now I think the normal view is to see cause and effect as simply a matter of an arbitrary cross-section or a line across this continuum. No particular line is privileged; it's just a matter of what you happen to find salient or interesting at the time. The whole idea of "causal powers" is redundant. And then he goes on to say that receptive properties are connections? What does that mean? How can receptive causal properties be connections?

 That is really the key to the system. A simple model would have individual elements each with its own effective and receptive facets, but as I understand it, Rosenberg prefers to see receptive properties as properties which connect individual elements. The complexes so formed have both receptive and effective properties and constitute "natural individuals". Causal relations between individuals and complexes help to impose constraints on the possible states of the component individuals, with the system tending towards higher levels of determinateness where possible. I must admit that the apparatus proposed here is rather complex and the motivation for some of the details is not always clear to me, so I might be misrepresenting the theory. It seems to be Rosenberg's idea that phenomenal experience is the ultimate substrate or "carrier" of physics itself. Moreover, the concatenation of simple elements into higher level complexes eventually gives rise to true consciousness: the Consciousness Hypothesis tells us that "Each individual consciousness carries the nomic content of a cognitively structured, high-level natural individual. Conscious experience is experience of he total constraint structure active in the receptive field of an individual."

 Frankly, that seems to me just an over-developed and unduly obscure version of a Higher Order theory. Causation looks to me like a primitive - one of those basic concepts you can't analyse. When you try to do it, the primitive keeps popping up in your explanation, reducing it to circularity. Don't you think that happens here? These individuals which get grouped together - they are imposing constraints on each other, but doesn't that mean they are

causing each other to be one way and not another? Yet we are supposed to be below the level of cause and effect here - we're meant to be explaining how causes work!

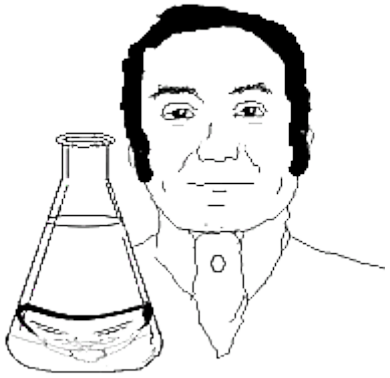
The real killer is this. Rosenberg set out to explain qualia, but at the end of the day it seems to me your real qualophile would say: yes, that's all very interesting, Gregg - thing is, I can imagine all of that happening without my actually experiencing the real redness of red. I don't see anything in your theory which actually catches the vivid reality of subjective experience. Now of course, in my eyes all talk of qualia is so much hot air, but I don't see why that would be any less plausible than the case for qualia was in the first place.

 I think the impression of circularity arises from your using an unduly loose sense of "cause", and I don't see how anyone could read a theory about relations between subjective experiences without seeing how it relates to qualia.

I don't buy the theory completely myself, as a matter of fact, but to me it's a very welcome piece of radical new thinking, and unlike some others, this is a book I intend to read again.

Dr Jekyll's Theory of Consciousness

30 March 2005



One issue raised by Gregg Rosenberg is that of Multiple Personality Disorder, or Dissociative Identity Disorder (DID), as it is now known. He uses it merely to support his case for the conceivability of co-existing but distinct identities, but it is a large, interesting, and unclear subject in itself; and the idea of multiple identities fits very nicely with theories of consciousness based on the "pandemonium" model, in which a number of agents or "demons" compete to determine the contents of consciousness.

DID has always been a controversial subject. Although there is an extensive literature and there have been a huge number of cases, it is still quite possible to disbelieve in it more or less completely. Actually, that requires some qualification. It really depends on what you think the phenomenon of multiple personalities amounts to. I see three main possibilities:

- the patients in question really have more than one separate person operating independently within a single brain;
- they merely believe they have more than one personality;
- they systematically maintain an elaborate pretence of having more than one personality.

Even the last of the alternatives constitutes an interesting psychological phenomenon, whose existence it would be hard to deny. Of course there is also a fourth, psychologically uninteresting category, consisting of those who systematically maintain the appearance of multiple personalities, not out of some obscure internal motivation, but in order to evade punishment, perpetrate fraud, or for other culpable but entirely normal motives. It is possible, prima facie, that there are real examples of cases falling into each of the four categories. People being the ambiguous creatures they are, the borderline between the second and third categories may be blurred, with patients who sincerely believe in their own multiplicity but occasionally help the evidence along a bit: and philosophers like Daniel Dennett , in the pandemonic tradition, would probably suggest that the distinction between the first and second category is not as sharp as we might have thought.

As it happens, Dennett, with Nicholas Humphrey, undertook an investigation of the DID phenomenon back in the late eighties, attending the Fifth International Conference on Multiple Personality/Dissociative States. An article for the New York Review of Books was apparently deemed not sceptical enough for publication there; eventually published elsewhere, it was included in Dennett's 1998 collection *'Brainchildren'*.

It's easy enough to see why Dennett might find DID an intriguing possibility: on his view the content of consciousness is the result of competition between multiple drafts; why shouldn't alternative drafters occasionally seize control and produce deviant versions of their own? But Dennett and Humphrey did not paint an uncritical picture of the conference. They noted the campaigning zeal which many of the delegates displayed, and the competitive boasts made by therapists about the performances of their patients. Some of the claims made across the dinner table apparently included accounts which any dispassionate scientist would find hard to accept: patients whose eyes would change colour when a different personality took charge; a

woman who was sterilised under one personality but became pregnant while controlled by another. Dennett and Humphrey noted that delegates whose research seemed unimpeachable were nevertheless tolerant towards colleagues whose claims and methods seemed less credible and remarked that they themselves found they almost automatically went along with the claimed multiplicity when talking to patients who switched from one persona to another in mid-conversation.

Ultimately, they considered that DID was a real clinical syndrome - though it was not clear how much of a role the therapists might play in shaping it or even bringing it about. This concern - that the doctors might be prompting patients to display dissociative symptoms - is a recurring one in this area. They noted that the number of cases was rising steeply - 200 up until 1980, 1,000 in treatment by 1984, 4,000 at the time of publication (there may have been ten times that number before the end of the century), and asked - where was all the multiplicity in earlier periods? In fact, DID has a slightly longer history than Dennett and Humphrey seem to have realised. Until the great upsurge in the late twentieth century, the number of cases seems to have been highest between the mid-nineteenth century and the first two decades of the twentieth. Both the beginning and the end of this earlier phase are interesting.

It has been plausibly suggested that the idea of multiple personality only emerged in the early nineteenth century when demonic possession (real demons in this instance) ceased to be a phenomenon which educated opinion (especially that of doctors) could regard as scientifically tenable. It is certainly true that possession by devils, dybbuks, and the like shows some marked similarities with DID, although there are perhaps differences, too. Modern DID patients often seem to slip from one personality to another relatively easily - even unnoticeably - while my impression is that diabolical possession was a much more traumatic business and less easily revoked. Possibly this reflects the change in social attitude: whereas the victim of possession was in for probable exorcism at best and possibly burning at the stake, DID patients, happily, face no such dangers and perhaps don't need to struggle so hard.

Prima facie, of course, it's quite possible that historical cases of possession were in fact DID; in principle it's equally possible that DID is simply a new form of possession by more subtle and mendacious demons. Perhaps in a less repressed and more secular society the devils don't have to struggle so hard, either. Joking apart, I think there can be little doubt that moving from an inexplicable spiritual phenomenon to one which was in principle susceptible to medical intervention represented an important advance: it also allowed the pervasive sense of evil attached to possession to be put on one side.

This sense of evil is very explicitly set out in the fictional work which, like it or not, hovers in the background of all these discussions. The Strange Case of Dr Jekyll and Mr Hyde first appeared in 1886 - which means it cannot be regarded as the source of the whole phenomenon, neat as that theory might be. Dr Jekyll believes all men are dual in nature: in fact, foreseeing the eventual advent of theorists like Dennett he says:

"*I hazard the guess that man will be ultimately known for a mere polity of multifarious, incongruous and independent denizens.*"

It so happens, however, that Jekyll's own transformation takes place along moral lines. Since he happens to be impelled by wicked motives at the moment of taking his potion, Mr Hyde is the purely evil facet of his personality: a relatively small part of the composite Jekyll, he is also

physically smaller (so much for a mere change of eye colour!) but gradually threatens to take over (not entirely, or not unambiguously, against Jekyll's will).

Given such deep roots and such a potent literary expression, one might wonder why the incidence of multiple personality ever faltered: but in fact it did fall during the early part of the twentieth century. One possible factor is a growth in awareness around that time of the need for therapists to exercise care in the way they handled patients, leading to fewer cases in which the disorder was evoked rather than discovered. Perhaps more significant is the emergence of new theoretical approaches; Freudianism, for example, offered an alternative approach to the analysis of mental illness which probably led practitioners away from diagnosing dissociation quite so widely. Particularly significant was the growth of schizophrenia as a diagnosis.

That, of course, opens up a new range of issues. The cause of schizophrenia is notoriously a subject on which many diverse theories have been proposed; but the definition of the condition also remains a matter of controversy: it has even been suggested that the term 'schizophrenia' does not, in fact correspond to any consistent and distinct psychological condition.

The word 'schizophrenia' was first introduced by Bleuler as a substitute for the misleading term 'dementia praecox': unfortunately the new word proved to have misleading aspects, too. In Bleuler's view the essence of the thing was abnormality in the pattern of associations, which is presumably why he chose a name which literally means 'split mind'. He did not mean multiple personality, but the derivation of the name somehow took a deep hold in the popular imagination. In films, jokes, and comic strips, schizophrenia remains synonymous with having a 'split personality' in spite of regular and repeated explanations to the contrary. It may well be that many of the patients who presented cases of DID during the second half of the twentieth century were under the impression (consciously or otherwise) that they were displaying the typical symptoms of schizophrenia. The misconception that 'split personality', under the guise of schizophrenia, was in fact one of the most common mental illnesses may actually have facilitated the growing incidence of reported DID.

These days, DID is regarded as just one form of a whole spectrum of dissociative disorders. One widely held view is that these arise from traumatic experiences undergone by the patient: the mind wishes so fervently to be elsewhere during these unpleasant episodes that it disowns the experiences, and may then attribute them to another, separate personality. Frequently the traumatic experiences in question take the form of child abuse: and some cases have been linked to alleged satanic ritual, bringing an unwelcome and probably unhelpful reminder of the devils once thought to be responsible for possession. I'm ill-qualified to have an opinion, but I don't find this theory appealing, in spite of the empirical evidence in its favour: it seems to resemble too closely the simplistic psychology of popular novels and films, in which people are generally 'driven mad' by some tragic or painful event. Even if true, the theory seems excessively teleological (the mind dissociates because it wants to) and offers little in the way of real explanation of the mechanisms which must presumably be involved.

So is DID itself imaginary, or at least iatrogenic (caused by doctors)? I think it is very unlikely that anyone actually has multiple persons functioning independently within them in anything approaching a literal sense. If anyone were to have dual personality, it would surely be those commisurotomy cases whose brains have effectively been split into two halves: on the contrary, however: such patients function with a high degree of unity and normality, and only carefully constructed experiments reveal the effects of the operation. Of course, there are big differences between physical bisection and psychological dissociation, but at the very least the split-brain

experiments demonstrate how robust and persistent the unity of the human mind really is. This makes the pandemonium model an implausible one in my eyes but does not in itself prove that views like Dennett's are wrong: a multiple drafts machine would surely need an exceptionally strong and resilient editor-in-chief, anyway.

Although DID may not be literally a matter of multiple persons, it's a little too easy to dismiss the whole phenomenon as the result of copy-cat imitation, the example of Dr Jekyll, and popular clichés about schizophrenia. The tendency to believe in the compellingly real presence of apparently non-existent people seems to be a deeply rooted human tendency - think of the imaginary friends children often acquire. I don't know of any fully satisfying answer as to why this should be, but some possible contributing factors include the notable predisposition of the human brain towards recognising people: the same bias which leads us to spot faces in random patterns. A second factor is that, while consciousness may be unified, it is certainly complex in some respects - as witnessed by the human capacity to carry out complex tasks unconsciously, and indeed, to talk to oneself. Perhaps the tendency to conjure up other persons also has something to do with the cosmic loneliness which according to Walter J Freeman , results from our cognitive isolation. Finally, perhaps, some role is played by the fact that like Dr Jekyll, we find the idea of another self seductive: even - perhaps especially - a wicked other self.

Boxing With Qualia

10 April 2005

The sub-title of Jeffrey Gray's book *'Consciousness - creeping up on the hard problem'* signals an attractive modesty about his achievement. Elsewhere he describes himself as ducking and weaving round the problem of qualia like a boxer, putting in a jab here and there and dancing away again, picking up points where he can without looking for a knockout. His approach is distinctive, using well-known elements in slightly unexpected ways, so that at times it's almost as though he's using judo moves in his boxing ring. I'm not sure his approach always works, but I think his strategy entitles him to a large degree of latitude over some of the issues, especially the philosophical ones, since he declares that though his book recognises the need to deal with all the issues, it is not a philosophical treatise.

One case where his philosophical stance seems dubious to me appears very near the beginning. We do not, he declares trenchantly, see the real world; only an illusion constructed in our heads. Far from having direct experience of an external world, we only perceive a mental model, full of things like colour and music which don't really exist in the world (other than as complex and rather arbitrary properties based on the differential reflectance of surfaces or vibrations in the air). Gray recognises the danger of this argument - if we never perceive the external world, what reason can we have to believe in it at all? He carefully affirms on several occasions that there really is an objective external world which has something to do with our perceptions. But I don't think he really realises the gravity of the problem. When I think about a cow, I think about a real cow, not a model cow in my head. One of the properties of thought is that I just have the power to do this. If Gray agrees with this, I think he ultimately has to abandon his view; if he disagrees, then the status of his assertions about the reality of the external world becomes problematic - can he even *talk* about the real external world?

A second pillar of Gray's view also seems debatable, if not to quite the same extent. This is the idea that all conscious experience suffers from a kind of time-lag. Gray takes Libet's conclusions (that our decisions are made before we even become aware of them) pretty much at face value, and he also quotes the case of tennis, where the stroke, it seems, often has to be played before there has been sufficient time for conscious perception of the ball. The idea of a systematic delay is undoubtedly tenable, but personally I suspect that the appearance of a delay arises merely from a failure to distinguish between decisions consciously made, and decisions which have had time to become the objects of conscious examination themselves.

Anyway, these two dubious propositions go to support a theory which in itself is not at all unreasonable. Gray has unconscious processes putting together a model for the examination of the conscious faculty while they (the unconscious systems) get on with actually playing the tennis, driving the car, or whatever. These unconscious systems seem so proficient at looking after all the detailed business of life that one wonders why they bother keeping consciousness in the picture at all - a question Gray goes on to address. There must, he says, be some Darwinian justification for consciousness; and in fact, he believes it serves the purpose of 'late error detection'. Normally, our behaviour is under unconscious control, but the model is used to extrapolate future situations: where unexpected elements appear or a process takes an

unexpected route, the matter is referred to consciousness where there is the opportunity for a second look at the circumstances.

The idea that something along these lines does happen is quite persuasive. Just as our eyes automatically jump to look at anything surprising, our attention focuses on the unexpected and the irregular. Gray also makes the claim that, rather paradoxically, consciousness also tends to identify features of the environment which remain constant: the unconscious processes operate on a moment-to-moment basis, but things tend to hang around in consciousness for an appreciable period. As a result, consciousness permits the identification of those common referents in the real world which we need in order to communicate with each other.

Any system which has a mental faculty perceiving a model is vulnerable to a charge of homuncularism or infinite regress: if perceiving the outside world requires a model which is then perceived by consciousness, doesn't the perception by consciousness require a second model perceived by an inner consciousness, and so on? Gray, justifiably, I think, rebuts this accusation as it applies to his own theory. The kinds of perception he is talking about are sufficiently different, and the helpful role of the model is clear enough to avoid any further regress.

I think the general scheme is plausible, but whether it accounts for more than one limited, attentional aspect of consciousness seems more dubious. On Gray's model, consciousness is in thrall to unconsciousness, constantly having its attention directed to potential problems. In reality, most of the time consciousness directs attention wherever it wishes and often goes on operating in complete detachment from current unconscious perceptions, whether problematic or routine.

Worse, while such error-handling processes may well occur, the theory does not explain why they are conscious, let alone why or even how they involve qualia. What has phenomenal experience got to do with it? Much of Gray's discussion of qualia after this has the air of shadow-boxing: the real qualia have escaped somehow. He argues, for example, that qualia can have real causal effects; but on his account it seems hard to understand why anyone should ever have supposed there was any difficulty about it: instead of the ineffable phenomenal qualia of other discussions, we seem to be talking about fairly straightforward pieces of data.

It must not be thought, however, that Gray sees it this way: in fact he presents a case against any functionalist analysis of qualia. One reason to reject functionalism is that Searle's anti-computationalist arguments strike Gray as convincing; but his main argument is a novel one. According to functionalism, he says, qualia are identical with appropriate brain functions. It follows that qualia and functions are in a one-to-one relationship: one function, one quale and vice versa. But synaesthetes get the same qualia from two quite different mental functions - from seeing a colour and from hearing a number, for example. This is no voluntary process, but one which the synaesthetes are powerless to dispense with even when it becomes a nuisance.

I think the functionalists have at least two escape routes from this argument. I think moderate functionalists need not actually identify qualia with functions: they might simply believe that qualia arise from functions (or indeed, they might just deny the existence of qualia altogether). Alternatively, it could be that the qualia *are* functions, and that seeing colour and reading a number (for the synaesthete) are two distinct functions which nevertheless both trigger the same qualic function. I don't see why that should be impossible.

Gray himself ultimately takes the respectable but limited view that qualia must indeed be reducible to some (non-functional) properties of brains: we just don't know at the moment which or how. He has some kind words to say about what he calls the Penroff (Penrose-Hameroff) theory, though stopping well short of endorsing it.

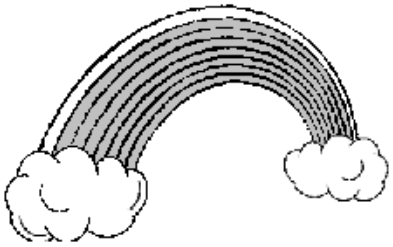
There also seems to me to be an air of shadow-boxing about Gray's treatment of intentionality . He rejects the idea that intentionality is an essential component of qualia, so it is not essential to his main project, but he does present an account of it. I think the canonical approach to intentionality (if there is one) would be through a discussion of that amiable couple Des and Bel (Desires and Beliefs), but Gray's approach is entirely to do with perception. While not illegitimate in itself, this does, I think, lead him up a blind alley. Based on a model attributable to Harnad, he suggests that the ground level is 'iconic representation': the simple impression made on the organism by the object perceived. The second level of the ascent is categorisation, in which the impressions are grouped together. And, well, that's about it. I think this is unsatisfactory in more than one way. Both 'iconic' and 'category' are treacherous terms which need very careful handling and explanation and remain insufficiently clear in this brief account. More fatally, the discussion doesn't take us as far as Gray seems to think. At most he establishes a basis on which intentionality might possibly be built: the claim that the account has arrived at 'full intentionality' just does not seem to be justified.

The foregoing presents a rather negative view of some of Gray's arguments, but fairness requires an acknowledgment that there is a great deal of fascinating and illuminating discussion in the book, on neurological and other topics, which we have not touched on here. In his concluding remarks, Gray says that some of his conclusions were not foreseen when he started writing, and he hopes some of the reader's preconceptions have also been disturbed. No theory, he concludes, is currently up to the mark, but some of the right questions are now being asked. That is a process to which he has certainly contributed.

The Octarine Argument

20 April 2005


One thing that particularly impressed me about Jeffrey Gray was his well-informed scepticism about colour. This must be a problem for people like you. According to you what people take to be colour qualia are actually just reality: the real qualities or properties of real objects in the real external world. If I can demonstrate that colours are not *out there* - that they're in your head, or in some Platonic world, or some phenomenal realm - you'll have to re-think, won't you?



You realise, of course, that colour is not a property of physical objects themselves, but comes from the light they reflect or emit? You probably suppose, however, that each colour corresponds directly to the wavelength of the light in question. Not so. The actual wavelength varies tremendously according to the light falling on the object; what colour you see depends to a great extent, not on the wavelength of the light coming from the object you're actually looking at, but from a complicated comparison with the wavelengths coming from nearby objects. This helps to ensure that (within limits) perceived colour stays constant in spite of varying light conditions: but if you isolate a coloured surface from the kind of clues provided by comparison with other surfaces, a colour can't be assigned to it at all - it just looks white.


I understand perfectly well that the colour vision system is complex, and that the brain infers colours from the available evidence through unconscious and largely involuntary processes. So what? At the end of the day, colour *is* a property of objects
- Gray himself concedes that colour vision is a fallible but practical guide to surface reflectance. Colour is an interesting property of real objects - if it were imaginary, as you seem to think, how come the colour of objects is so consistent and such a good guide to various physical and chemical properties? Colour vision is useful enough to have evolved separately several times in different species - if it wasn't telling us something real, how could it be so useful?


Yes, yes - colour *corresponds* with real properties. But colour itself is not in the external physical world. It's like those cases where a picture from a telescope or a microscope is given colours artificially to help show up the different substances. The colours are not real, but they do help you understand more about what you're seeing. It turns out that all colour is more or less like that -yes, it's helpful, but that doesn't mean it directly reflects the physical reality.

After all, we only actually sense three different wavelengths. Our eyes have cells which respond with different degrees of enthusiasm to three different colours: all the others are inferred from the ratios between the three. A good thing, too, in a way, because otherwise colour photography would be a lot more difficult. As it is, we can give the appearance of millions of colours just by different mixtures of red, green and blue.

If you think about it, though, this means that the way we see things is actually wrong. To us, a combination of red wavelengths and green wavelengths looks the same as pure yellow wavelengths - but it *isn't* the same. Compare sound, another wavelength phenomenon. If I play middle C on the piano and the E above it at the same time, it doesn't sound like D. But with vision, the wavelengths just get averaged out. Part of the reason, of course, is that your ears don't attempt to assign sounds to a precise location in space. They can afford to have one accurate sensor (a little hair-like thing) for lots of different wavelengths. But your eyes need to assign colour to a precise spot - so they can't crowd hundreds of different cells into that one small area of the retina.

All the same, it is possible to do a bit better, and some animals detect more than three different wavelengths. Pigeons can do four, for example, so they actually see more colours than we do. Now how do you cope with that? According to you, colours are real: but in that case pigeons somehow have a different reality?


I don't claim that our senses are perfect; pigeons may see colour more accurately, certainly, but I absolutely deny that they see more colours, except in the trivial sense of being able to discriminate finer gradations. They see the same spectrum we do; just more clearly.

Ask yourself this: if we only sense three wavelengths, why do we see more than three colours? Of course, there can be an indefinite number of different mixtures of the three, but I think it's obvious that the spectrum we see contains more than that - the full ROYGBIV seven. You rightly pont out that a mixture of red and green light looks the same as pure yellow light: but if we can only sense the red and the green, why don't they both look like mixtures? But they don't. They look like yellow: and yellow doesn't look at all like any mixture or combination. The spectrum is real, and our senses infer that reality and present it to us even though they can't detect it comprehensively.


That sounds borderline loony to me. What about people who are red-green colour-blind? If our senses somehow manage to infer this Platonically real spectrum in spite of inadequate evidence, surely they should be able to see the full spectrum, too? And as for ROYGBIV - you're actually going to defend the separate existence of *indigo*? Everyone knows that Newton stuck that one in just to make sure the number of colours was the mystical value seven.


Red-green colour-blind people do indeed see the full spectrum. You don't imagine that they look at a rainbow and find that the red or the green stripe is missing? They see both red and green - they just have difficulty telling which is which.

I don't know about indigo particularly, but I think Newton was right about assigning seven colours to the spectrum. Perhaps this is another case, like gravity's action at a distance, where his mystical leanings allowed him to entertain ideas that orthodox scientists had an allergic reaction to.

Consider the analogy with sound. Both are wave phenomena. Sound comes in octaves, each with seven distinct notes. When you get to the eighth note, it is a higher or lower version of the first one - C above or below middle C, say. Now the basis of the octave is a doubling, or

halving of the frequency, and that's why the two notes sound the same. If a wave comes along and twitches one of those little hairs you were talking about eight times a second, it matches up with a wave which twitches it four times or sixteen times. The other notes on the scale come from slightly more complex ratios. The Greeks knew all about this. Now the visible spectrum is just the right length to provide seven notes out of an octave of light. And that's what we get.

 So you maintain that the colours we see are the *only possible* colours? What about octarine?

 What's that - an eight-legged citrus fruit?

 Octarine is Terry's Pratchett's imaginary eighth colour. Obviously magic is involved in his case, but there has to be an octarine, doesn't there? The mere conceivability of another colour shows that the spectrum is not an absolute reality. It seems to me that, just as we can always encounter a completely new smell, there would always be scope for a new colour, if our eyes were able to develop new responses the way our nose presumably can. But I don't even need to rely on conceivability. Some insects can see ultraviolet light, for example, and some snakes can see infrared. They must assign to those wavelengths colours which we can't see, mustn't they?

 Not at all. Look at the way the spectrum forms a closed circle. If we extended it downwards below red, we should simply get another, lower, violet. Now I grant you that the 'lowerness' would have to expressed in some way - possibly as 'warmth'. The colours of the visible spectrum are differentiated in terms of warmth, so perhaps the lower violet would appear distinctly warmer than the one we're used to (great scope for interior decorators...).

I repeat, the spectrum is a reality. You can call it a mathematical reality if that helps, but it's real. If we saw colour the way we hear pitch, all this would be obvious. But the fact that we can't see colour harmonies or more than a single octave of colours means there's never been any scope for a genius to come along and produce a regularised interpretation of the spectrum, the way J.S.Bach did for the musical scale.

 I'm simply amazed by all this. But at least it is nice to hear you indulging in speculative phenomenology.

The Future of Intelligence

1 May 2005

It's a sign of the times, perhaps, that Jeff Hawkins is not a computationalist. Creator of the PalmPilot and the Treo smart phone, he approaches the brain from a technical angle. The philosophy doesn't concern him unduly. He describes a conversation about qualia and reports his strategy: "Well, I don't see anything special going on, so if you do, I must be a zombie. But hey - that's OK." You would expect someone with these leanings to be a diehard AI merchant, but not a bit of it. While he believes it's feasible to build artificial brains that work the way human ones do, he thinks they would be profoundly different from the computers we know and love, and expects entirely new kinds of silicon chip will be needed.

Actually, it's a bit of a misrepresentation to suggest his background is one of computers: he has also spent a good deal of time on neurology and in fact founded the Redwood Neuroscience Institute for the purpose. My impression is that the more you know about neurology, the less appealing the computer analogy becomes; but it seems that Jeff has always been of the view that the only way to reproduce the human brain was by close study of the methods the brain itself uses.

All this, and his own theories of how the brain works, are set out in *'On Intelligence',* a book he wrote with Sandra Blakeslee. (Sandra Blakeslee is a journalist who has specialised in this area: she previously helped Ramachandran write his notable book *'Phantoms in the Brain'.* I should think her own thoughts would be well worth reading if she ever publishes them.) The book has a fully-developed website of its own here .

Hawkins, as I say, has little time for good old-fashioned AI: in fact, he endorses Searle's persuasive story about the Chinese Room. Much more to his taste are the approaches taken by neural networkers and connectionists, though they too, in his eyes, don't go far enough in modelling the actual neurology of the brain: he laments the fact that much of the connectionist effort has been expended on relatively simple three-layer networks. Both the AI and connectionist projects suffer the fatal flaw of focussing on behaviour: they see the brain in input-output terms, almost in the way the behaviourists did. While behaviour is obviously important, the real key to the mystery of intelligence is what goes on inside, as is obvious when you consider how much of our most important mental activity goes on while we are at rest, and without any direct connection to our behaviour. Imagining, remembering, and even planning can all go on quite effectively while we are sitting quietly in an armchair. Remembering, as it happens, turns out to be especially important in Hawkins' theory.

Hawkins is inspired by Vernon Mountcastle's suggestion that all regions of the neocortex are essentially doing the same thing. Although there are some differences in the detail of the physical structure, and some regions have been shown by long-established research to be associated with particular senses or functions, the proposal is that at a fundamental level the processes are all the same. This is not such a strange idea as it sounds: it is supported by the remarkable flexibility of the cortex: although particular functions typically end up in particular

places, a variety of evidence shows that in cases of injury or abnormal development, different regions can swap functions around quite readily.

If all the different parts of the neocortex are doing the same thing, then what is that thing? Hawkins proposes that the answer is *prediction*. To summarise with extreme brevity, the neocortex takes the incoherent swirl of data coming in from our senses and constructs invariant representations of objects in the world. These stable, invariant patterns allow us to recognise the same object from varying views or from partial information, and because they include patterns with a temporal element, they allow us by analogy to predict what is likely to happen next. In fact, only the things which fail to match up with our ongoing prediction of what is about to happen get referred upwards to higher, full-conscious regions for handling. Of course, memory is just as much a fundamental part of this process as prediction, and in fact what Hawkins is getting at is a wider power of extrapolation over time or space, rather than prediction understood narrowly.

This all seems pretty sensible, though not uncontroversial; moreover it is accompanied by a fairly full attempt at explaining in neurological terms how the process is actually implemented. Hawkins explains the layered structure of the cortex and the columns which cross it vertically. The key problem is how on earth the brain manages to construct invariant representations. Seeing the same person never involves exactly the same set of sensory inputs - we see different parts of them from different angles in different lights, yet the brain almost effortlessly recognises the same person.

One view is that the invariant representation is achieved only at the end of a complex process: Hawkins suggests that it is invariant representations all the way. Each layer of cortex deals with different representations and feeds its results up to the next, so that the invariant representation of lines and shapes builds up through hierarchical development to the representation of such complex objects as people at the highest level (actually Hawkins suggests that the memory-forming hippocampus should be seen as in a sense the highest level of cortex). This too seems sensible, but I wonder whether it is the whole story. There is an implication that every possible experience is already implicitly wired into the brain. There may be senses in which this is unobjectionably true: but I find it difficult to assess whether the unimaginably complex wiring of the cortex is really up to the indefinitely huge task involved. I certainly don't think it's clear that it isn't, but of all the steps in the account, this seems to me the one most in need of reinforcement.

Hawkins goes on to consider the prospects for intelligent machines (I suppose we can't refer to them as artificial intelligence); in fact, he offers a rallying cry to young engineers: this is the moment to start work if you want to get in on the next big wave! He thinks that the limitations on connectivity suffered by silicon chips, compared with neurons, might be offset by the comparatively high speed of wires: the silicon neurons can share connections regulated like a phone network and still achieve similar results. It sounds reasonable, though there is an evident danger of underestimating the subtlety and complexity of real neurons - a danger Hawkins should particularly be sensitive to. In discussing the possible applications of such artificial brains, he does seem at one point to slip back towards a computer-oriented view. One of his proposed applications is the control of cars (The attraction of this is doubtful, I should have thought, because his artificial minds will learn like humans and surely therefore be as fallible as biological chauffeurs?). He suggests that once a suitable artificial brain has learnt the art of driving, the factory will be able to run off any number of perfect copies. I'm not so sure - are these new chips going to be programmable? It is a leading property of computers that they can

be put into any desired state directly, but that is not true of all machines. Some things have to be assembled in a certain intricate order: the desired state is only reachable through an appropriate chain of prior states. I have a feeling the Hawkins chips would each have to learn driving individually.

All in all, though, Hawkins' views seem unusually practical and promising. He wraps up by offering eleven detailed and falsifiable predictions about the neurology of the human brain: these will be interesting to watch.

Cycosaurus Breaks Free

12 May 2005

A recent New Scientist piece announced that within the next few months, CYC would be unleashed on the Internet. The CYC project (pronounced 'psych', but derived from 'en **cyc**lopedia') dates back to an earlier, more optimistic era of Artificial Intelligence; it was the most confident of those based on 'brute force' - the assumption that the problems could be conquered by huge databases and massive computing power, rather than requiring some revolutionary new insight. The survival of the project for more than twenty years, and its emergence now seems as surprising to me (and almost as impressive) as the sudden appearance of a Tyrannosaur on 21st century streets. But perhaps that just shows how wrong I've been.

The problem confronted by CYC is the same, hydra-headed monster which in different ways has blocked the advance of most forms of AI - the problem of dealing with the chaotic complexity of real-life, real-world problems. Computers do well at solving simple problems and carrying out tasks in which every contingency can be nailed down in advance, but that is rarely possible in the wider world.

A classic example often used in explaining flow-charts is the making of a cup of tea: we're asked to break down the job into stages (get the pot, fill the kettle), consider the correct order in which tasks should be carried out, and build in some elementary branching (is the kettle full?). This is fine so long as we're dealing with imaginary teapots in a grossly simplified world: but actual real-world tea-making is a jungle of unforeseen problems and bear-traps which no computer could cope with. Just recognising the teapot is a formidable task: teapots don't have to be made of any particular material or be any particular size: they can be any shape and even their topological properties are not guaranteed - they don't*have* to have a tubular spout or a handle. Is that a teapot we're heading for, or might it be a kettle, a coffee-pot, a jug, a picture or a mirror? Can we put down the paper we're holding, or will we lose track of it forever? When we move the kettle, should we note its new position, and must we also note that the cups, the walls, the table and the floor have not moved? Have the cups, as a matter of fact, fallen on the floor because we failed to anticipate that moving the kettle laterally would shove them off the table?

Humans move through this infinite sea of troubles fairly effortlessly. Computers always seem to run into one or other of a set of related problems. Either the vast number of possible permutations in reality gives rise to a *combinatorial explosion* which exhausts the speed and capacity of even the most powerful computer, or the frame problem rears its ugly head, either in the classic form of being unable to keep track of what hasn't changed, or in the looser philosophical form, where we are unable to pick out the important contingencies from the infinite range of possible things we might think about.

In some expositions, these problems, especially the last one, seem so daunting that it begins to seem obvious that intelligence is actually impossible. But humans pull it off somehow. It seems that human beings have access to a vast body of knowledge about the world - a *background*, a *common sense* - which just tells them what is likely, or relevant, or reasonable, in any given context. One view is that this relies on a special faculty, a kind of intuition which we need to

understand better before we can replicate it. There are plenty of theories about where this special faculty might come from, but they tend to be unfriendly to old-fashioned conceptions of AI. Another, pragmatic, approach has been to set up *frames* or *scripts* , which supply the computer with assumptions about particular limited contexts. This approach yielded some good results in the early days, but it suffers limitations almost by definition; we're still essentially talking about toy worlds.

It might be that until you achieve a certain minimum quantity of factual 'common-sense' knowledge, a level well above anything yet realised by most projects, the thing won't really fly at all: if so, then CYC might prove to have been the only viable strategy. It certainly represents the most serious attempt to tackle the whole problem head on - to assemble a significant fraction of the entire corpus of human everyday common-sense knowledge, and use it to give a computer the same kind of background understanding that human beings automatically possess. Even CYC, of course, is not supposed to know *everything* immediately. A key part of the approach is an ability to generate new truths from its stock of information, using a system which is ultimately based on predicate calculus

There's no denying, I think, that Doug Lenat, the founder and guiding spirit of the project, has pulled off a remarkable achievement in keeping such a large project on the road. When the original funding dried up, he set up Cycorp to carry the project forward: as progress was made various cut-down versions of the software, notably OpenCyc, have been made available to others to study and work on. You would expect that as the data accumulated, the software would gradually find it harder and harder to cope, but Lenat says that it actually gets easier for CYC to draw its own conclusions as time goes on, rather than being spoon-fed. This, it seems, is the rationale behind the creation of a direct internet-based interface with CYC; it will soon be so well-informed that it will be able to talk directly to people and gather its own information. I don't think this interaction is likely to take the form of chat-bot style conversations, but even a yes/no question facility would be interesting.

Can it be true, though? Are we really on the verge of breaking through to human-style comprehension, with all that that implies, and has the trick been pulled off by good old-fashioned AI and massive computing power after all? I'm afraid I think it is highly unlikely. Neither an encyclopedia nor predicate calculus seem to me to work in ways which are even remotely like the human brain.

Now that may seem rather sweeping. The brain does store a huge number of facts, and a well-informed person can indeed supply information, just as an encyclopedia does. Equally, the human brain does do predicate calculus - indeed, if it didn't, there wouldn't be any predicate calculus. But in both cases, these are very specialised forms of thought. An encyclopedia records facts in a fixed, explicit form: if you want to run off the same piece of text, well and good: if you want it even slightly reinterpreted, a formidable processing task awaits you. If I ask you, 'Do dogs have tails?', you answer without difficulty, and if it contains a suitable statement, an encyclopedia could answer just as readily. If I ask the question in a different form - 'Are tails usually part of a dog's anatomy?' - you answer just as easily. You don't have to ask yourself whether dogs have tails and then reason from there to the required conclusion. The encyclopedia, however, probably lacks the exact formulation required by the second question and a different process has to be invoked. If, moreover, I ask whether dogs wear beards, the question is easy for you (barring fancy breeds, cartoon dogs, and so on), but it will be a non-

trivial task, perhaps an impossible one, to deduce the answer from the information in an encyclopedia

Deduction, after all, is a relatively weak and limited way of generating new conclusions. There is a strong impression about all of formal logic that you only get out what you put in, whereas normal human thought encompasses a range of creative, intuitive and insightful processes which far outrun pure logic. Human beings do predicate calculus the way they do tight-rope walking; by artificially restricting and then bringing to a new level of perfection, a special form of one of their more general abilities. To try to build a model of human cognition on the basis of predicate calculus is like trying to replicate human locomotion by stringing ropes everywhere.

Of course, CYC has been around long enough to address some of these problems, and gain a sophistication which earlier conceptions did not have. For one thing it seems that instead of tackling the whole world in one bite, CYC has a series of theories about particular fields or aspects of reality. It includes, apparently, a full statement of the Linnaean system to help it recognise and classify animals, for example. The fact that different theories are applied in different realms allows CYC to tolerate contradictory or inconsistent beliefs to some degree: lack of this ability is one of the leading weaknesses of AI systems based on classical logic, a problem which others have tried to address with non-monotonic logics or other systems.

But this too raises an objection. Common sense is supposed to be the opposite of theory, not an example of it. This mistake was exemplified by the "Naive Physics Manifesto" issued by Patrick Hayes back in the 70s: he saw the normal human understanding of physics as based, not on Einsteinian or even Newtonian theory, but a kind of secret, subtly mistaken theory which held that things have an inbuilt tendency to stop moving, and so on. The attempt to reconstruct this naive theory so that it could be built into artificial intelligences proved an interesting failure, and I believe the real reason was simply that normal understanding is not theoretical in nature at all.

Academics have a natural tendency to value theory rather highly, and they are prone to mistake the tools which consciousness uses to investigate reality (encyclopedias, formal logic, computers) with the mechanisms which sustain consciousness (and even reality) itself. Perhaps CYC is an example of this rather clerkish tendency. If so, the bad news comes with a possible silver lining. CYC might never be truly intelligent in the way human beings are, but it might still turn out to be a formidable tool.

I must admit that the idea of the old monster breaking out and rampaging triumphantly through the streets is one I am reluctant to abandon altogether.

Chess

22 May 2005

Computers playing chess must be about their most celebrated achievement in the fifty-odd years since Alan Turing launched the modern quest for artificial intelligence. In a way this is odd: chess is only a game, after all - what does it matter? One reason must be the high social status of chess and its association with esoteric intellectual prowess; no-one would have been impressed to the same extent by successful poker-playing robots, in spite of the formidable challenges that raises.

But a more important reason is probably the way chess had previously been used as the paradigmatic example of something computers couldn't do. If we go back as far as the 1960s, a standard account would explain that computers were very quick at maths, but incapable of certain other tasks. The only way they could play chess, for example, was by exhaustively considering all possible moves and their consequences: an astronomical task which no machine, even in theory, could possibly deal with. It was usual to illustrate the way permutations grew rapidly out of control by telling the old legend about the inventor of chess. According to the story, he presented the game to the king, who was so delighted he offered the inventor any reward he chose. The inventor, with apparent modesty, asked for a quantity of wheat: one grain on the first square of the chessboard, two on the next, and so on. Of course, it turns out that the quantity of wheat involved by the time we reach the square 64 is in fact colossal, far beyond the capacity of any mere kingdom to supply: it has been variously described as several hundred times the world's current annual production; a quantity which would require a barn about 25 miles square and 1000 feet high; or a quantity which, counted out grain by grain, would take about 584 billion years to deliver. Such is the power of geometric progression. The inventor got his head cut off instead.

Those 60s writers, however, underestimated the scope for progress in two important respects. First, even computers don't have to exhaust all possible moves: more sophisticated algorithms allow more efficient problem-solving. Second, they didn't take account of Moore's Law. First stated in 1965, this was originally about the number of transistors you could get into an integrated circuit, but it is now generally interpreted as the wider view that computing power itself increases geometrically, doubling periodically. It was as though the king in the story had come up with a magic grain of wheat which produced two identical offspring, each of which went on to do the same, and so on.

For some time it was an interesting question as to whether chess would eventually be beaten by clever new programming which let the computer play in a more human style, or by the simple advance of brute-force number-crunching power. In practice, it became increasingly clear that some combination of the two was inevitably going to work one day. The sceptical view died hard though; Hubert Dreyfus notoriously disparaged the chess-playing potential of computers in his 1972 book '*What computers can't do*', and has been the object of gleeful mockery more or less ever since. He denies ever saying that computers would never be able to play chess: rather, he remarked that at the time of writing they still couldn't even play at a good amateur level. Dreyfus has a sophisticated theory about the limitations of AI, along Heideggerian lines: essentially he denies the possibility of formalising general-purpose human cognition, which in his view really consists of a complex mixture of habits, skills, and other non-theoretical capacities: as he also points out, this stance does not naturally imply that computers would perform badly in a

formalised micro-world like chess. As the century wore on, chess-playing machines and programs filled the shops, and the best programs began to creep up even on chess masters.

The argument was effectively settled, of course, by the famous victory of 'Deep Blue' over Kasparov in 1997. Actually, there are still some arguments which sceptics can deploy. Deep Blue was carefully adjusted during the match on an ongoing basis by human experts to match Kasparov's play and the strategic situation: this parallels the way human champions, including Kasparov, can take advice from a a support team during intervals, but perhaps a more thorough test of the silicon side would have merely allowed Deep Blue to interface with other computers. The presence of human 'advisors' in the process opens the way to an accusation that the match was more a matter of 'play by wire' than genuinely autonomous computation. In the final analysis, though, does it matter? Do computers have to able to beat the world champion before we accept that they can play chess?

Should we, in any case, be impressed by the chess prowess of contemporary computers? I think we should, for two reasons: one fairly evident, the other slightly obscure but rather more alarming.

The victory of Deep Blue was presented as very much a matter of brute force computing power - we were encouraged to think that Kasparov had been beaten by a machine, not by a program. This is obviously an over-simplification, but the conquest of chess does represent a victory of sorts for mere processing power, and this has implications for other fields. Computer translation, for example, which has really only achieved the most modest levels of success so far. The underlying problem is that some parts of translation depend on understanding the meaning of the text, which computers can't do. But the task is also surely amenable to brute force in theory. If the program has enough information about alternative translations, and a long enough list of contextual clues to pick up on, the level of error will eventually fall below that of human translators (who, after all, sometimes misunderstand the text themselves). In principle, the 'list' required is vast, possibly even infinite, but if chess is any guide, that need not be an obstacle. Chess has still not been exhausted, and probably never will be: it turned out to be enough to deal with a salient sub-set of cases. It might well turn out that a similar subset of things people are likely to have written would be enough for practically efficient translation. Once translation is out of the way, a seriously effective performance on the Turing Test begins to look possible at last. Passing the Turing test unambiguously isn't really a sign of consciousness (certainly not for a machine which we knew was working on brute force principles) but it would certainly open up a distinct new phase of the argument.

The second point springs from the discovery, implicit in the success of brutal computing, that there are distinctively different ways of 'thinking about' chess than the way the human brain does it; it turns out that in some cases these may reveal things the unaided human brain would never have found. The most remarkable achievements in this respect are perhaps brute-force analyses of endgames rather than general chess playing ability. By exhaustively analysing the possibilities, computers have shown that many positions which were previously considered to be inevitable draws can in fact be won, in some cases by extraordinarily long drawn out sequences of moves. In the case where a king and two bishops faces a king and a knight, it had been chess orthodoxy since the mid-nineteenth century that if the 'knight' side could obtain a certain position, it could draw. This proves to be entirely wrong: in fact, the 'bishops' side has a win in almost all cases. Some of the endgame 'strategies' found by computers are capable of being learnt by human beings, but this particular case is an example of a strategy which is so meandering and bizarre that in spite of considerable efforts, it seems to be beyond the power of

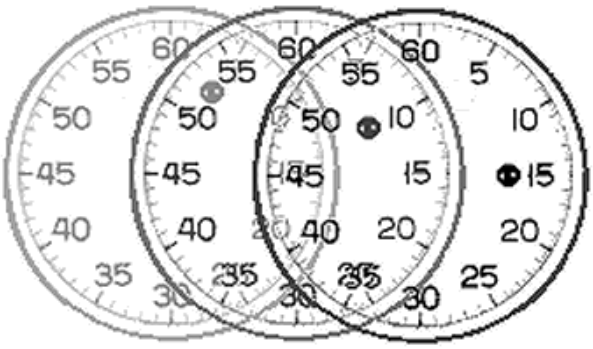
even a specialist to understand or learn it. Presumably the number of different tactical contingencies which need to be held in mind at the same time simply exceed the capacity of the human mind, sometimes said to be limited to about seven items.

Personally, I find this rather scary. This kind of discovery can perhaps be related to the computer-inspired finding that chaotic results can emerge from the repeated application of relatively simple equations, quite contrary to human intuition, and more loosely to the emergence of mathematical proofs by exhaustive computer examination of all possible cases: proofs which we have to accept but cannot really ever understand.

In short, it suggests that human cognition is actually rather patchy: the gaps in our comprehension of the world in general may actually be quite large, but we don't notice them exactly because they consist of the kind of thing we have trouble recognising. Of course, if Colin McGinn is right, consciousness itself falls into one of these blind spots.

Libet's Short Delay

2 June 2005



The experiments carried out by Benjamin Libet into the timing of conscious awareness have provoked, and go on provoking, a vast amount of discussion. His own theory of consciousness as a kind of field has received somewhat less attention; and the strange brain-cutting experiment he proposed to test it seems likely to remain unperformed for the foreseeable future. A large number of papers and discussions have been published: in 2004, Libet finally summarised his own account in the book *'Mind Time'*.

Libet's early research was actually intended to explore what the minimum stimulus giving rise to a conscious sensation might be. He had an enviable opportunity to study the response of the brain to direct stimulation (using trains of electrical pulses) through the help of a friendly neurosurgeon and the co-operation of a series of patients, who remained conscious and able to report their sensations throughout the experiment. There are several ways of varying electrical stimuli, of course, but a curious fact emerged: whatever the voltage or frequency of the pulses, the stimulus had to persist for about 500 milliseconds before the subject became consciously aware of it. Actually, this is not quite true: above a certain level of voltage, the interval decreased, but the current involved was by then well above anything likely to occur in the brain normally.

The result was unexpected, because it had already been established that stimuli applied to the skin, rather than directly to the brain, could be detected consciously even if they were much shorter than 500 milliseconds. Libet was able to demonstrate that although the stimulus to the skin might be brief, it was still the case that the resultant brain activity had to persist for 500 milliseconds before the subject became consciously aware of it. Where a patient was anaesthetised, the initial brain response to a stimulus was the same as for a fully conscious subject, but it failed to continue for the required period: moreover, a stimulus applied directly to the brain 500 milliseconds after one applied to the skin could cancel (or in some circumstances, enhance) it.

Now, it isn't surprising that there should be some delay between an event, and our becoming aware of it: indeed, if the normal process of cause and effect is to be sustained, the event has to precede the awareness it causes. If we were passive spectators of the world, simply watching the way we might watch a film, the constant delay would be irrelevant -we should never notice that we were half a second behind reality. But we also respond to events, and here a delay is highly relevant and noticeable. The really surprising thing, therefore, was the length of the delay which seemed to be involved. 500 milliseconds - half a second - is a noticeable period of time, and it is evident that human beings often respond to events far more quickly than that. If we had to wait half a second before responding to events, we should never be able to play a good game of tennis, and we should be dangerous (or extremely cautious) drivers.

The answer appeared to be that our unconscious responses are far quicker than our conscious ones. A stimulus applied to the skin produces an 'evoked potential' or EP in the brain within tens

of milliseconds, and that seems to be enough for it to register unconsciously but effectively. A series of experiments have shown that we register unconsciously a whole host of things which may influence our response to events but which never cross the threshold into consciousness. Among other evidence, Libet quotes experiments which show that a conditioned response - a blink - can be created to events which the subject is never actually conscious of. The remarkable phenomenon of blindsight might perhaps be seen as a related case.

That still leaves us with a considerable problem. If the foregoing is true, we ought to be aware of it, surely? On the tennis court we would find to our surprise that we returned serve competently before we actually saw the ball, and certainly without thinking about where in the opposite court we might want to put it. Our conscious and unconscious behaviour would be strangely unsynchronised. Libet's hypothesis was that conscious awareness is subjectively referred backwards in time. We consciously perceive the stimulus as occuring at the same moment it registers unconsciously, even though it doesn't in fact enter our awareness until it has persisted for half a second. Subjectively we backdate it to match the EP at the beginning rather than the end of the 500 millisecond span.

Libet was able to provide some direct evidence through experiments which, instead of comparing skin stimuli with direct stimulation of the cortex, instead made a comparison with stimulation of the medial lemniscus, part of the incoming neural pathway. Each pulse delivered here generates its own EP, but the sensation is nevertheless referred back to the time of the first in the train. Backdating remains controversial, however. Perhaps the sensations actually enter conscious awareness immediately, and the half-second delay merely allows time for them to become reportable, or fixed in short term memory? Perhaps we are merely dealing with the difference between being aware of the stimulus, and being aware that we are aware of it? Libet contends that awareness and memory, especially declarative, explicit memory, are different and independent phenomena.

There is a further problem to solve, in any case. The idea of back-reference allows Libet to eliminate the chronological inconsistencies which threatened to develop in perception, but that leaves our perceptions at odds with our decision-making. We now seem to perceive the ball before we perceive the answering stroke of our tennis racquet, but since both seem to occur half a second before the *actual* moment of conscious awareness, we should, bizarrely, be seeing ourselves begin to play the appropriate stroke just before we have consciously decided which it is.

Confronted with this problem, Libet decided that one more step in the argument was necessary: the perceived time at which we make a decision must also be subjectively referred back by 500 milliseconds. Unlikely as it seems, and contrary to our own impression, we must have made our decisions slightly before we actually become aware of them.

Devising a further experiment to test this hypothesis was challenging on two counts. First, how can you tell when a decision has been made? In the case of perception, we know when the stimulus occurs, and can track when the initial EP appears in the brain, but the only symptom of a decision seems to be the resulting action.

However, research carried out in 1965 by Kornhuber and Deecke showed that when subjects were asked to move their wrist of fingers at a moment of their choosing, the act was preceded by a measurable electrical change in the brain. This 'Readiness Potential' or RP appeared about 800 milliseconds in advance of the act and seemed to be a clear indication that the intention to

act had formed. It therefore seemed that the RP might be used experimentally as a marker for the decision.

The second problem lay in how to measure the moment at which the subject became aware of having made the decision. If the subject had to press a button or give some other sign, the careful timing would be obscured by the time taken in doing so - the experiment would be measuring *two* hand movements and their associated decisions! Libet's solution was to arrange an oscilloscope so that a bright dot circled around a kind of clock face every 2.56 seconds (no doubt these days a PC would be used and the circuit time set to a more rounded figure - but this was 1977). The subjects were simply asked to note the position of the dot at the moment when they became aware of having decided to move their wrists. This allowed the timing to be reported accurately while taking the time required for the report out of the equation.

Of course, the results confirmed the hypothesis: the RP appeared some 500 milliseconds before the reported awareness of a decision to move.

That sorted out Libet's account in purely empirical terms: in philosophical terms the problems were only beginning. The research (subsequently repeated and corroborated by others) seemed to provide a scientific proof that free will was a delusion. How could we consider ourselves responsible for decisions we were not even aware of until after they had been made? Some would have been happy to see the demise of free will, but Libet himself was not ready to let it go so easily. Although the subject's decision to move occurred too early for it to have been initiated by conscious thought, there was still - just - a window of opportunity in which conscious awareness might conceivably veto the move. This window lasts, on Libet's account, no more than about 100 milliseconds. Experimental proof is difficult. Libet has conducted experiments in which the subjects were asked to form an intention to move and then veto it at the last moment: apparently an RP appeared and then dissipated, but the weirdness of the mental gymnastics required of the subject seem to leave an element of doubt about the process. Is it possible to decide to move at a random moment while simultaneously holding on to the belief that you will not, in fact, execute the movement?

Interestingly, Libet has a moral argument here. He rightly points out that free will is important partly because it underpins the idea of moral responsibility: but morality, he suggests, is mainly a matter of vetoing things we have a built-in tendency to do. This is a faintly depressing prospect, and I think Libet underestimates the difficulty of discriminating between doing and not-doing. If I veto my momentary desire to kill you, that surely is morally good: but what if I repress the automatic tendency to grab your hand when you are about to fall off the cliff?

There are several avenues of attack against Libet's other conclusions, of course. Is the RP *really* a signal that a decision has been made? If I make a decision about my insurance policy, does an RP appear, or is it just wrist movements that cause RPs? The circumstances of both Libet's experiments and the earlier ones by Kornhuber and Deecke are rather strange: they require the subject to get into a frame of mind where they are ready to make a decision any moment. Might not the RP merely signal a quickening of attention, rather than a moment of decision?

Libet believes that by timing the moment of awareness through his oscilloscope arrangement, he eliminated the need for the subject to spend any time on reporting the moment of awareness: but isn't it possible that we need a certain amount of time just in order to report the awareness to ourselves? Awareness of the decision you have made is one thing,

being aware of that awareness is another - which might well be thought to require some further time to develop.

Personally I also doubt whether it is necessary to reduce free will to a veto system - 'free won't' as it has been described. Libet often seems to take it for granted that every free act is preceded by a specific act of will: but that isn't really the case. Often the conscious mind sets a general plan, on which we then act more or less automatically. A tennis player has thought in general terms about how to play the next stroke long before the need for actual action: drivers have a kind of running rule in the back of their mind to the effect that if something suddenly appears in front of them, they hit the brake. Free will operates at this higher level, with all our actions being managed in detail by unconscious processes. I don't have to think about where I want to hit the ball at the very moment of decision in order to control my game of tennis any more than I have to think separately about each of the individual muscles I am implicitly proposing to contract.

Libet has another string to his bow, however, inasmuch as his general theory of consciousness is also designed to offer a foothold for free will, along with the subjective experience which in his view sustains it: free decisions are conscious decisions, and conscious decisions require subjective awareness. He proposes a conscious mental field (CMF) intended to account for the various mental phenomena which don't appear to be natural consequences of the firing of neurons. Like other proponents of field theories, Libet believes that the prescence of such a field helps explain how the diverse activity of the brain is bound together into a single, unified conscious experience. Unlike Susan Pockett or Johnjoe McFadden , however, he does not see this field as a straightforward physical phenomenon. Libet, who says that when young he was convinced of the truth of determinist materialism, no longer believes that conscious mental activity is explainable by or reducible to, neuronal activity, although it certainly requires it. He is not, at the same time, advocating any kind of dualism or spiritual theory: instead (and rather obscurely, I think) it seems that his proposed CMF is an emergent phenomenon: something that arises from the combination of active neurons but amounts to something distinctively more than, and different from, the sum of brain activity.

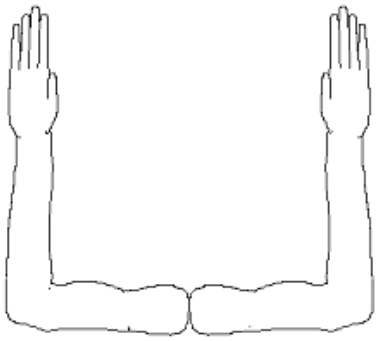
In order to verify the theory, Libet has proposed a rather alarming experiment. If a small slab of living brain in a suitable area of the cortex, could be cut off from its neural connections while being left with a suitable blood supply, he believes it would remain capable of producing subjective experience, mediated by the CMF, when appropriately stimulated. (Actually the slab would have to have some minimal neural input to keep it 'awake', but this could apparently be managed without spoiling the experiment.) The existence and role of the CMF would thereby be demonstrated - though I imagine there would still be considerable resistance to the idea and a host of objections to the experiment.

The immediate theoretical objection which springs to mind is to ask why the brain goes to such trouble to make neural connections if it can do its normal job without them: Libet's answer is that it can't do the 'normal jobs' of information processing, memory, emotional response and all the rest: it's just the subjective aspect which arises from the CMF.

Of course, there are also practical difficulties: you cannot cut out slabs of living human brain purely for experimental purposes. Libet has identified a method and a category of patients in whom the experiment could theoretically be performed innocuously, namely those who are already going to have an epileptic focus surgically removed. Very few patients of this kind exist, and Libet has had no success in interesting the surgeons in these cases in facilitating his experiment.

Egyptian Consciousness

20 June 2005



There is a tendency in discussions of consciousness to talk as though nobody had any particular ideas about the mind before Descartes 'invented' the soul and 'introduced' dualism into European thought. When we consider the idea of the soul at all in connection with consciousness, we generally adopt a kind of neo-Cartesian conception of it: a simplified Christian view of the kind which was accurately, if insultingly, summed up by Gilbert Ryle with the phrase 'the ghost in the machine'. This isn't really fair even to the Christian tradition, which ranges from extreme mysticism, like that of Meister Eckhart (who thought human souls superior to those of angels and spoke of passing on to a plane beyond God) to the resolute materialism of Hobbes, (who believed that even God was corporeal). Non-Christian traditions get even shorter shrift.

This neglect of ideas from older sources is is not actually altogether unreasonable: you wouldn't go searching through ancient Chinese or Hittite texts for the solution to a modern engineering problem, so why should you expect the same sources to provide anything of more than historical interest on consciousness? The trouble is, the mental clichés which remain unchallenged at the back of our minds may make it difficult to think in new ways about the problem, confining us to the same old narrow tracks we have so often traversed before; trying to get our minds round some of the varied ideas of the past may help us see a wider range of possibilities. The scale of what we may be missing is made clear by the sheer difficulty of grasping properly other conceptions of the mind: even the Greeks, in spite of their colossal influence on the way we think, are very difficult to get right in this respect. The word *psyche* and its derivatives are ubiquitous; but in spite of having been translated as 'soul' for hundreds of years, it doesn’t really mean that, or 'mind' either; in fact there isn't a single English word which exactly equates to it. It comes from a word meaning 'breath’ and means something like 'a source of animation'. One way of illustrating the difference is to note that in English the question of whether animals have souls, or minds, is a reasonable one which different people, for different animals, would answer in different ways. To an ancient Greek, asking whether animals have psyches would have seemed absurd, approximately equivalent to asking whether animals are alive.

Of course, it's never easy to be sure that we have the meaning of an ancient term right, especially since many of them had different meanings at different times and in different mouths; the Greeks are helpful in this respect in providing a quantity of explicit argument about the significance of particular concepts (though their passion for originality exacerbates the problem of establishing a single authoritative meaning).

When it comes to the case of the Ancient Egyptians, we are in a much worse position. The Egyptians were practical people, not generally much interested in abstract matters (not even maths, beyond what was necessary for effective building projects)

and with relatively little taste for dissent and argument. Our picture of their views, moreover, has had to be constructed from whatever hieroglyphic texts remain.

One thing which is abundantly clear is that the ancient Egyptian conception of the soul and personal identity was not a simple matter: we're dealing with at least five distinct elements in addition to the body, the *khat* . The first noteworthy point is how little there is of obvious philosophical interest about the khat. The Egyptians went to enormous lengths to mummify and preserve it, and it was considered capable of reanimation; it had a role as host to other elements of the person, but it seems fair to say that the Egyptians essentially thought of the khat as an inert vessel in much the same way as later dualists did. They would have thought that Descartes had missed many of the important points about the soul, and they would have thought the importance he attached to the pineal gland as the site of the vital interaction between soul and body was ridiculous - the Egyptians regarded the entire brain as little more than packing for the skull - but they would have been quite at home with his general position on the body.

The first of the soul-like elements in the Egyptian system is the *ka* . The word 'ka' is often said to be untranslatable, which raises an interesting question. If the word is untranslatable, and is known to us only from hieroglyphics - our translations of hieroglyphics - how do we know what it means at all? 'Ka' has often, in practice, been translated as 'double', though it seems this is misleading. It is sometimes represented as a smaller image of the individual concerned, sometimes as a pair of arms reaching upwards, either detached or on top of the head of the individual concerned. It has been suggested that the word is a kind of pun on a similar Egyptian word meaning 'you', which presumably would carry the suggestion of the ka being the essential you. It is a life-force, not altogether unlike the psyche, and like the psyche it was not unique to human beings: animals and plants had their own ka, too. Apparently Egyptians sometimes spoke of having a snack or a drink for your ka, rather in the way that people might speak of refreshing the inner man, or of keeping your spirits up, perhaps. The ka was not just a quantity of life-energy, however - your ka is specific to you, bestowed at birth.

I think it's plausible to see the ka as being, like the psyche, the thing that gives you the power of moving. Surrounded as we are by machines which move in complex and controlled ways, we have lost any sense that the power of spontaneous movement is particularly remarkable; but in ancient times it must have seemed obvious that this was a quality which distinguished people and animals from everything else, and required some special explanation

The ka really needs to be considered together with the other main soul element, the ba. The ba is represented by a human-headed bird (again, a pun may be involved, though for us the wings might evoke angels, the dove-like paraclete of the Christian trinity, or ancient Greek harpies). Although the ka is specific to each individual, it is really the ba that contains all the characteristics of the individual, especially the personality. Whereas animals had kas, the ba was something human beings had in common with gods. Although both were to some degree dependent on the khat, the ba was tied to it rather more closely; the ka could wander off even during sleep and after death was the first to move on to a new afterlife. When the person successfully negotiated the judgement which awaited them after death, the ba eventually joined the ka and they merged, forming the final immortal stage of the person, the *akh*.

So were any of these elements of the personality considered to be conscious? Not the khat, except insofar as it was animated by the ka and ba, but both of those appear to have had the power of independent thought and action in some degree, and the akh presumably conjoined

the two somehow. It's impossible to be sure, but I would guess that rationality and explicit thought were properties of the ba, while the ka had the kind of thinking without words and vivid phenomenal experience which we might attribute to our nearer relatives among the animals. At the risk of oversimplifying, they could perhaps be seen as the conscious and subconscious minds; from the point of view of contemporary philosophy of mind, it's very tempting to attribute qualia to the ka and explicit cognition to the ba.

So far the system hangs together quite neatly, but we also need to take on board two other elements which were clearly linked with the ka and ba, but which are much more mysterious. These are the *khaibit*, the shadow, and the *ren*, the name. The name, given at birth (at about the same time as the ka, presumably) was considered to have a constitutive role: without it, in any case, the individual could not survive. On a few occasions the Egyptians actually attempted to erase all instances of a dead person's name, in the belief that this would destroy them in the afterlife. The name, in fact, apparently enjoyed a kind of afterlife of its own, distinct from that of the akh.

Clearly there is an unobjectionable sense in which someone's name constitutes them; gives them their agreed place in the social system which in some respects is the essential ground for personality and identity. And of course, it can be said metaphorically that a person's name sticks with them through life and lives on after death. But it seems clear that the Egyptians had something more metaphysical, if not magical, in mind; more reminiscent of the tremendous power ascribed to names by the Christian legend of the Adamitic language, the true original language of the Garden of Eden which captured the essential reality of the things it named. One might well ask what the ren actually named; the ka, the ba, the akh, the khat, or some composite which ought, unsettlingly, to include the ren itself.

The role and capacities of the shadow are even more obscure: it was linked with the ba and was said to move at great speed, though the ka seems to have been less constrained. It is interesting (again from a contemporary point of view) to notice that both represent forms of intentionality. The name is the means by which the world, or at least other people, shape and target the individual; the shadow is a prime example of the impact the person, in turn, has on the world. If the ba is the seat of cognition, a strong link with the intentionality of the khaibit might be natural, and the alleged speed of the shadow might be the speed of thought itself, which can address distant things instantly.

That completes the six essential elements most normally described; but I think it is also worth mentioning the ib, the heart. I said above that the ka and ba united only after judgement; but it was not any of the foregoing elements which faced that judgement. The ib, in fact, was weighed against feathers, and if it failed the test was eaten up, resulting in permanent extinction for the individual.

It is tempting to see the ib as the seat of moral responsibility within the Egyptian scheme, and make the judgment after death comparable to the Christian one. However, if the ba is the seat of reason and deliberation, it would naturally be the entity held accountable for the person's acts. My guess is that the ib was in fact put in the scales to establish its weight, not its worth; not to assess whether the person was good, but whether they were morally substantial. A person whose acts had been evil but strongly willed and considered might pass such a test while a harmlessly passive one might fail. It would follow that the ib was not conscious, but more like an attribute of the individual.

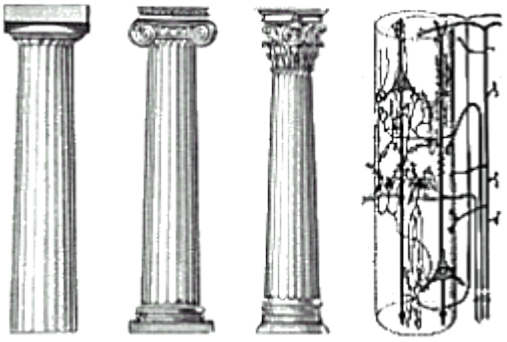
This does not by any means exhaust the Egyptian vocabulary of personhood, but I think it is enough to be going on with; I am already in danger of trying to bring together systematically views which may not actually have been part of a fully coherent system. A twenty-first century person, after all, might well believe simultaneously in the Christian soul, ghosts, the Freudian subconscious, and the multiple personalities of DID, without ever bothering to work out exactly how the different conceptions fitted together, and the chances are the Egyptians were similarly unworried by the odd complexity of their views.

However that may be, and whatever the correct interpretation of the Egyptian system is, I think it is at least clear first, that their ideas were not at all simple, and second, that if they were not Cartesian dualists, it was chiefly because a mere two-part theory could not have done justice to the polyverse multiplicity of the Egyptian soul.

Blue Brain

3 July 2005

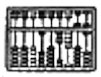

Well, it looks like the beginning of the end of the argument so far as AI is concerned. A true brain simulation is on the way: it's only a matter of time.



There's always been a kind of last resort argument for those of us who think the brain must ultimately be capable of simulation on a computer. OK, you don't accept that any of the existing theories is any good; OK, you think when computers play chess or whatever, they're actually doing it in a completely different way. But at the end of the day, the brain is just another physical object. If it comes to it, we can just simulate *the whole thing*. We can model every single neuron in a real brain. What then? We don't even need to understand how the brain works; if we copy it exactly, it will work, just as if we copied all the parts of an old-fashioned watch without having any ideas about horology. If we do that, we just will have consciousness functioning in a computer - there really is no denying it short of some hopeless retreat into dualism or whatever. Now you've always said that such a model was impossible, not just in practice, which seemed obvious, but also in principle. Guess what? Somebody's actually going to do it.

The Blue Brain project, (a joint venture of the Brain Mind Institute, EPFL, Switzerland and IBM, USA) has the eventual aim of applying terrific computer power to the simulation of an entire brain. It's not going to happen overnight, but it's *going to happen*.


Hang on a minute. Aren't you exaggerating? So far as I can see, all they plan to do is simulate one of the columnar structures in the cortex. That's a tiny fraction of a whole brain.


That's the initial stage. It's a significant step forward in itself, but the ultimate aim is a total brain simulation. What I'd like to know is what you think makes this impossible? What's going to stop them, given enough raw computing power? If you accept that the brain obeys the same laws of physics as the rest of the world, I just don't see what's to stop a simulation. Really, you're just betting against the progress of technology. That's always been a bad idea. You're putting yourself in the same camp as those people who said heavier-than-air flight was impossible in principle.


I really need to know more about the project, I think. A lot depends on the level of the proposed simulation. Your view has always been that there is a relatively high level of interpretation on which we can find a working functional analysis of the brain - we can identify the essential working parts without having to go down into fine detail, more or less the way we can identify the way the cogs in a clock work without having to worry too much about what they're made of. But it ain't necessarily so. My guess is that these people are going to use a standard model neuron and put a whole lot of them

together. But there's a tendency to under-rate the neuron. It isn't a simple switch. We know now that there are lots of different neurotransmitters that work in different subtle ways and are influenced by the chemical environment in all sorts of ways. These fine details may well be important, and it's not unlikely that to get a working simulation you'd actually have to model each individual neuron down to the molecular level, which is just beyond any conceivable technology. Even then, you might well find that you needed to model the surrounding glial cells at a similar level of detail. People tend to think glial cells are just sort of stuffing that fills up the gaps between neurons, but there's been some research which suggests their role in getting neurons ready to fire again may be quite significant. The complexity of all this is mind-boggling.

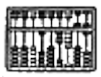

I very much doubt if it's as bad as that. You have to remember that however sophisticated the behaviour of neurons may be, they ultimately only do one thing: fire - or fail to fire. I actually think that modelling all the vast number of connections is going to be more of a challenge than worrying about neurotransmitters or glial cells, but in any case, get this - they *are* modelling on a molecular level.


What! I... I actually don't think I believe that.


Well, it actually isn't completely clear what the whole-brain project will eventually be like, but certainly some of the existing work is at a molecular level. And in the current columnar project, I believe each processor is supposed to handle only a few neurons - plenty of capacity for really fine detail, I should have thought.


Hmm. Well, I think they're going to run into some problems. Actually, in some ways, doing an entire brain is a more hopeful project than doing part. I assume they picked on a column from the cortex because, as well as being an interesting structure which probably does have something to do with higher mental functions, it looks a bit like a component; a single unit that can be unplugged and run in isolation. But that isn't really true. Granted, the cortex has this columnar structure, but it also has a definite laminar structure - it comes in layers. These columns are strongly wired up to other structures as well as internally; simulating one might end up being as useful as simulating a single cog out of a clock. But even doing an entire cortex might not work; we know that other parts of the brain are essential to certain kinds of memory, and to emotion, for example. It might easily turn out that you need almost the whole brain before the simulation will really work. I believe people can get by fairly well with most of their cerebellum missing, as a matter of fact, but apart from that it may all be essential.

I have to ask - don't you suspect this is a just a marketing ploy on the part of IBM?


Oh, it clearly is, but so what? Doesn't mean it won't work. I see your point about the column in isolation, but I don't think anybody is naive enough to think it's an unpluggable component. It's just such an obviously salient structure that it's clearly worth studying.


The other thing is, whose brain are they going to be simulating? There are lots of detailed differences between the way one person's brain is wired, and another's. Now it's

quite likely that some of the properties of that pattern of wiring are essential to consciousness, while others are just individual variation. If you don't have some theoretical understanding already, how are you going to be able to tell which is which?

 You really are clutching at straws now, aren't you? I can see it must be traumatic when events force you to reconsider your whole intellectual outlook.

 The point I'm really making my way towards is that a simulation is a simulation, and reality is reality. Searle pointed out that you can simulate a thunderstorm on a computer, but nobody gets wet; surely it would be the same with a brain: you might simulate the way it worked, but you wouldn't get actual consciousness.

You see, it's important that the causal relations operating in an artificial intelligence are the same (in the relevant respects, whatever they may be) as the causal relations operating in a real brain. But if it's a computer simulation, they never will be. In the real brain, the firing of neuron A causes the firing of neuron B: in the computer, both events are caused by the instructions in a lot of code. It's the difference between someone who walks because of the muscles in their legs, and someone who walks because a puppeteer is pulling the relevant strings.

 No, that's false. It's more like this. Suppose you had a universal shape-mimicking device. Now you set some of these little machines to mimic the shape of cogs in a clock. When you put it together, it all works: now they're just simulated cogs, and internally they're doing all sorts of calculations and monitoring in order to keep the right form: but does that mean it isn't a clock? That it doesn't tell the time? Obviously not. The causal relations in the simulation are more complex, but they've been set up in such a way that the overall pattern is the same. Of course no-one gets wet from computer rain: all that means is that our brain simulation won't generate sloppy piles of white and grey gunk. It's not the physical nature of the thing that matters; it's the ability to perform in certain ways - yes, ultimately to have a coherent conversation, for example. You know this is true really - you don't recognise consciousness by the physical qualities of someone's brain, but by their behaviour - whether they answer you and what the quality of their answer is. This idea about some magic property of biological brains is an idea that would never have occurred to you if you hadn't needed a bolthole to hide in. When a real simulation comes along, that bolthole is going to become pretty uncomfortable.

Granny Neuron

23 July 2005

A little while ago the New Scientist reported research which apparently revived the discredited theory of the 'grandmother neuron'. The piece attributed this theory to Jerry Lettvin, and unfortunately this idea was picked up and widely echoed elsewhere. In fact, although I think Jerry Lettvin probably is the author of the *phrase* 'grandmother neuron' (it is also sometimes attributed to Horace Barlow), I believe he devised it as a rhetorical aid to his *rejection* of the idea. Alas, I dare say the meme that he endorsed the idea is now unstoppable. The grandmother neuron is actually one of those ideas that no-one really wants to advance, but which many people would like the opportunity of refuting. I suppose this makes the new research especially interesting, at least in principle.

What is a grandmother neuron? The idea is about the encoding of memories in the brain, and the idea is that one memory, or the memory of one thing, is represented by one neuron. There is, for example, a neuron that stands for your grandmother: that neuron is the only one which invariably fires when you think of your grandmother and never fires when you think of anything else. Other neurons fire in a similar way for your grandfather, your cousin, your house, an oak tree, and so on. There is a neatness about the idea which appeals, and some suggestive empirical support - we've known for some time that in experimental animals there are neurons that fire only in response to certain movements in the visual field, for example, and Wilder Penfield's classic research showed that electrodes applied to specific areas of the brain may evoke particular tunes or other remembered items.

But there are some problems. If your memory of your grandmother is constituted by a single neuron, it's rather precarious. Losing that neuron would suddenly render you unable to think of your grandmother. Intuitively, moreover, it seems odd to suppose that thinking of your grandmother is an on/off matter, like the presence of a 1 or a 0 in some grandmaternal register somewhere. Surely when you think of your grandmother you entertain a mental image of her, or recall her voice, her habits, or some significant events in which she was involved?

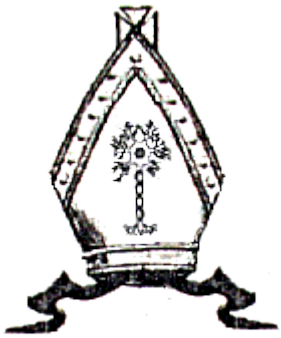
For these reasons, and because the research has generally suggested that many neurons are active when thinking about even simple things, it is almost universally agreed that mental representations are in some sense distributed: that to think of Granny requires the firing of a set or pattern of neurons. Some would suggest that the idea of Granny is built up out of a set of neurons representing various qualities and characteristics the old lady possesses - one set of neurons representing the shape of her face, one set of neurons which fire for relations, one that fires for all old ladies - but this seems to me almost as unsophisticated as the grandmother neuron idea itself. The truth is that we are on the fringes of the mystery of intentionality here, where there are few satisfactory answers.

What about the new research then? It seems the researchers identified a neuron which fired only in response to images of Jennifer Aniston; the images were different, so it must have been some general Aniston-ness (at least, of a visual kind) which caused the response. While this is not uninteresting, it seems to me to fall well short of reviving Granny Neuron. The researchers did not establish that no other neurons fired in response to Jennifer Aniston (nor, as a matter of

fact, that the same neuron would respond to Jennifer Aniston pictures on another occasion). Of course, as with many other experiments, this one was possible only because it could be fitted in alongside a therapeutic investigation, and its scope and nature were naturally constrained by the medical requirements of the exercise. But what exactly was established? If thinking of Jennifer Aniston involved the simultaneous firing of a distributed set of neurons, one would rather expect that there would also be at least one particular neuron which fired on every occasion. It might yet be that grandmother neurons exist, or that they don't; these experiments seem to sit equally comfortably with either hypothesis.

Pontifical Neurons

5 August 2005



Some, it seems, believe the neuron has a particularly special role, even more important than holding the memory of Granny. A paper in the Journal of Consciousness by JCW Edwards (mentioned by Steve Esser in a comment here) suggests that consciousness itself is actually a property of single neurons. The theory is clearly one of some stature; it apparently received favourable consideration from William James, and resembles the views of Leibniz – presumably the ones set out in the *Monadology* .

But why would anyone think that single neurons are independently conscious? In Edwards's eyes, this is the only likely solution to the notorious binding problem. The binding problem, as you may recall, is the question of how we manage to combine unsynchronised flows of information from our different senses into a smoothly unrolling, multi-modal conscious experience of reality. The suggestion here is that in order to achieve that unity of experience, the relevant bits of information all have to come together in one unit somewhere. A brain, or the cortex, or even a neuronal column, is just too big, it seems, but a single neuron would be just about right. The idea seems to be that a neuron is a real unit in a sense which more composite things, such as brains, can never be. At the same time, however, Edwards doesn't want neurons to be *too* strongly unified. A traditional view would be that neurons merely add together the total of inward impulses, and fire if that total is greater than a certain value, but Edwards believes they are capable of more complex behaviour.

Stated baldly, the idea that neurons are 'real' units sounds rather dubious - cells are pretty composite things themselves, surely? Edwards seeks to translate arguments from William James into terms more suitable to contemporary physics, but I'm not sure that adds anything to their underlying philosophical plausibility. There are other obvious objections; why would we have a head full of neurons if only one is needed for consciousness? When Edwards addresses this question directly, his answer seems highly unsatisfactory. He tells us that having a large brain is actually no big deal - the behaviour of small-brained bees is vastly more complex than that of cows living peacefully in a field. That seems a bit of a slur on bovine society, but in any case it seems to make the big brains even harder to explain. The real answer, implied elsewhere in the paper, seems to be that neurons have a kind of dual role: they do perform all the typical computational stuff, and that function is essential to their awareness of the world, but at the same time they each have a general, independent consciousness. It's as if they're each getting the full picture from the network around them, but at the same time unconsciously carying out some much simpler processing which contributes to that network.

So, to raise another obvious objection, what happens when the particular neuron which provides our consciousness dies – do we die with it? It was considerations of this kind which led James and others to speak facetiously of the supposed 'pontifical' neuron implied by the theory; but Edwards asserts that his system is democratic. A particular group of neurons, probably in the cortex, but perhaps in the thalamus, all have full consciousness. They all independently suppose that *they* are the person in whose head they reside. It follows that they are all having very similar experiences and thinking very similar thoughts - otherwise, surely, at least some of them, the ones not currently in direct control, would be disconcerted by the mismatch between their thoughts and the actions of their body.

This raises an interesting question - how do we count consciousnesses? If each neuron has its own consciousness, which seems from the inside to be the consciousness of a human being, and if each of these consciousnesses has virtually the same contents in other respects (to the extent that it doesn't matter which of them is actually in the driving seat), what reason have we got for calling the whole package a set of many consciousnesses rather than a single, perhaps slightly fuzzy, one? Presumably the only reason for assuming multiplicity of consciousness is the multiplicity of the corresponding neurons, but that in turn focuses attention on an unclear part of Edwards' theory - what does he suppose the relationship between consciousness and neuron to be?

I can see two probable answers. One is that the activity of the neuron is what 'gives rise' to consciousness; the other is that the consciousness has an actual physical location within or around the neuron. The first position is a difficult one for Edwards to take because - if I've interpreted him correctly - the whole network of neurons is required to give rise to consciousness, and it seems to follow that the whole network is the thing that 'has' the consciousness. The second position is tenable, but seems to me quite implausible. Why should consciousness have a location, any more than the number 5 has a location? We don't suppose that my consciousness of income tax has to be located near the tax office, so why should it be located near my head, either? The position of my eyes, ears, nose and mouth gives the impression that 'I' am located within my skull, but a virtual reality set-up can give the impression I am located on Planet Zarg. Consciousness itself doesn't seem to be the sort of thing that, in itself, has a location in space.

It may be that Edwards has a more subtle conception of the relationship between neuron and consciousness than I realise - but I think some clarification, at least, is required.

The real problem, anyway, is that the theory doesn't actually solve the binding problem at all. That was supposed to be its motivation, but Edwards' theory seems to *require* that the right impulses from different sensory channels arrive at the neuron (at each neuron, actually) simultaneously. But if we could arrange for that to happen, we should already have solved the binding problem: we'd already have found a way for the brain to pick out which sensory impressions should be put together. Slam-dunking them all into a single neuron seems more like an expression of our triumph than a means of achieving it; and indeed, if we had solved the basic problem, we might ask ourselves whether compressing the results into a single neuron really achieved anything.

Why Be Conscious?

19 August 2005

A good question. A paper by Lee Pierson and Monroe Trout asks, what is consciousness *for* ? If consciousness has evolved in human beings, it must have had some survival value, but what exactly was it? This reminds me of the way Wallace, the co-discoverer of evolution, eventually came round to the view that the theory he and Darwin had discovered could not account for human intelligence. Consider a simple hunter-gatherer, he said; nothing in his way of life requires the ability to do calculus or write symphonies - yet we have developed these supererogatory abilities.

Wallace was surely mistaken, but the question remains. Pierson and Trout's answer is, in a word volition. They believe consciousness gives humans (and other creatures - no species prejudice here) the ability to rise above the hard-wired responses of simpler creatures and manage their behaviour in a more strategically effective way. This is possible because consciousness confers the ability to direct one's attention, keeping it on an object for sustained periods, or for that matter diverting it to a new object on the fringes of awareness.

This is a pretty good answer, I think, though it doesn't strike me as outstandingly novel. We've heard before of how lower animals (and according to some even primates) are liable to have their attention dominated by what they can actually see here and now; the unfortunate chimp, for example, who finds it impossible to choose the smaller reward even though he has seen, and even seems to understand, that he is always given the reward he *doesn't* reach for.

Pierson and Trout range much more widely than that, however, and in doing so they sometimes become less convincing. Reading the paper is a bit like flying a plane above some low-lying cloud: in places you can see the ground unrolling quite nicely, but in others alarming peaks rear up unexpectedly without much warning. It may be that the theory suffers from being expounded in a short paper, and that a longer treatment would clear up some of these cloudy areas.

I'm not at all sure, to take a fairly fundamental example, what metaphysical status they are assigning to consciousness. They say very clearly that they regard it as an emergent property of neural activity, but they also seem to lean towards fields and other more novel explanations. The great advantage of emergence, surely, is that it gives you something unexpected without requiring new fields or other apparatus. There's also an awkwardness about saying that an emergent phenomenon has strong causal relations with the substrate it arises from - or at least, some elucidation is needed.

Another feature of the theory which seems to need more explanation is its anti-determinism. Pierson and Trout are very clear that the operation of volition is not deterministic. They repeat one dreadful argument against determinism. This goes as follows: according to determinists our beliefs are determined by our circumstances, but that means that the determinists' own views were simply dictated by fate, and they have therefore no grounds for saying that determinism is actually true: determinism undermines itself. A pretty weak argument: the fact that my views were caused by something hardly means that they are false: indeed, if they were caused by the

right kind of thing - relevant evidence and valid arguments - it might be a reason for thinking them true.

I suspect that Pierson and Trout just take it for granted that the sort of higher-level control they are proposing requires something non-computational; but that's far from clear - at the very least it requires some clearer supporting argument. The real problem is that it isn't at all clear how consciousness helps, either. How do subjective experiences allow us to direct our attention? For that matter, how does consciousness allow us to make decisions about which object to direct our attention towards?

I suspect that there might be interesting answers to some of these questions, and perhaps Pierson and Trout could give them - but they don't seem to have done so to me in the current paper.

Information

1 September 2005

'Information' is a slippery but seductive term. The concept of information is clearly tied up with many of the essential issues of consciousness, but it isn't easy to come up with a snappy definition which covers all the uses of the word. Paul Young says the word *"has no definition that establishes its identity as a physical phenomenon valid for all its uses".* His view, set out in his book '*The Nature of Information'* and now handily summarised in a shorter paper of the same name, seeks to elucidate the matter and open the way to a solution of many mysteries, consciousness among them. (You can see selections from the paper on his site; to buy the whole thing costs $20.)

Others have thought that an analysis of information might be the crucial place to crack the problem, of course. David Chalmers notably included some speculative reflections on information as the basis of consciousness in *The Conscious Mind*. I think it may be the very ambiguity of the concept, the way it seems to have a foot in both camps, which suggests it might be the elusive bridge between the physical and the mental.

But the same ambiguity involves a considerable danger; that of switching senses in mid-stream. Now whatever vagueness may be attached to the word in general usage, it does have at least one meaning which is clear and useful, namely the sense given to it within the information theory attributable to Claude Shannon. Shannon was concerned with the transmission of messages, and his theory allows useful calculations about quantities of information, transmission channels, compression, and related matters. But this is a rather specialised and limited sense; above all, it has nothing to do with semantics or meaning. In real life the messages sent down a wire generally have a meaning, of course, but so far as this kind of information theory is concerned, the meaning is irrelevant: the message is merely a specified set of binary digits. In Shannon's sense, information is everywhere, and the chance variations in the surface of a cliff carry just as much information as one which has an inscription carved into it. This is at odds with the more normal everyday sense, in which by contrast, meaningfulness is the essence of information. In this sense, we might say that data is just numbers: it only becomes *information* when it has an interpretation - when it means something.

The danger, when talking about consciousness, is apparent. One of the key mysteries of consciousness is meaningfulness or intentionality; the semantics which Searle insists you can't get from mere syntax. If we start our discussion with one sense of information, it seems clear that we are talking in simple physical terms; if we then, perhaps without even noticing it, slip into the other sense, we smuggle meaningfulness into the discussion without having given a proper explanation.

Young is too clear and careful to fall into the trap quite as easily as that, but some kind of evasion along these lines remains the most obvious threat to his account. He sets out the stages of his proposed progress, however, with exemplary clarity, as follows.

1. *Information, in every sense in which it is or could be used, both within and outside science, will be seen to be the exact equivalent of what generally is known as form*.

2. *Form will be redefined as a mass-energy or physical, rather than an abstract phenomenon.*
3. *The fundamental creative and control mechanisms of the universe will be seen to be embodied in its use of forms as information – i.e., all information is the processing of mass-energy forms.*
4. *All mental events, including cognition, feelings, volition, and all forms of consciousness, including consciousness of self, will be able to be defined as flows of mass-energy forms, and whatever it is we call mind will be seen as a wholly mass-energy, form-manipulating process.*
5. *The universe then can be defined as a mass-energy system that exercises its creative, control and communication functions by manipulating forms of itself – i.e., structures, patterns, shapes and arrangements or organizations of matter, many of which retain their identity through inward and outward flows of mass-energy, allowing us to see ourselves as forms of a self-organizing, self-controlled, mass-energy universe. As physicist, Victor Weisskopf wrote, "Nature, in the form of man, begins to recognize itself."*

Most of these proposed steps look ambitious to me, but perhaps stage 2 is the most challenging of all. Isn't form irreducibly abstract, by its very nature?

Form is, in any case, the key concept in Young's exposition, with faint echoes of Plato. Form is constituted by sets of relations between physical (or as Young would say, mass-energy) objects, and the flow of form, expressing it with extreme brevity, is information. There is a sense in which Young seems to want to have his cake and eat it; he doesn't want form to be an abstraction, but clearly, as he acknowledges, a form has to be divorced from any particular physical expression of it, and it's rather hard to see how it can remain a concrete fact while retaining such generality of application. There's a certain form which is common to the paper text of Pride and Prejudice, a magnetic tape recording of a reading, and a stream of electronic digits which encodes a video version, but the three examples have nothing physical in common so far as I can see.

That said, the account is generally a reasonable and persuasive one. As a point of view, or a perspective, it seems unexceptional; but does it amount to a theory? I gradually came to feel that I was being offered a plausible but general account of the physical correlates of mental processes, but that the vital and long-awaited moment when the lead of mere physics turned into the gold of conscious experience never really came. There are few statements in the paper I felt strongly moved to disagree with, but I did finally part company with the author over neurons. After describing the way neurons process their inputs, and giving a quotation from Theodore Holmes Bullock *("The neuron is like a miniature person... it speaks finally with one voice, which integrates all that went before.")* he declares that *"This activity by a single neuron is de facto concept formation - abstraction..."* Maybe it is, but I don't see how or why the activity of a single neuron can accomplish such a feat. I can see how sets of neurons in the visual system manage to respond only when presented with a line orientated a particular way, but does that amount to having the concept of a line? More seems to be required, though since I haven't read the book-length version I might, in fairness, be missing something.

It is also only fair to point out that the ambitions of the paper, though wide-ranging, are limited so far as consciousness is concerned: I don't think Young is professing to solve the "hard problem" of qualia (which he bypasses), only the supposedly easier problems of understanding and volition. It also seems that his target is the kind of consciousness shared by many other animals, rather than anything distinctively human. Nothing wrong with that, of course, and the

vision which motivates his account remains appealing, encapsulated in another quotation (the paper is liberally sprinkled with quotes from a variety of authors): *"Nature, in the form of humans, begins to recognise itself."*

The Loebner Prize

1 October 2005

The 2005 Loebner prize was awarded to Jabberwacky recently - or rather, to George, one of the virtual personalities supported by the Jabberwacky software. The Loebner prize, as you may know, puts chat-bot programs through a version of the Turing test. The judges have a series of on-line conversations, some with real humans, some with the chat-bot programs; the bot which most nearly persuades the judges that it is in fact a human being, gets the prize. There is a gold medal waiting for the first bot which actually passes as human, but there seems to be no immediate risk of it's being won.

I believe serious AI researchers have, on the whole, tended to stay away from the Loebner (it seems that in 1995 Marvin Minsky offered a "prize" of $100 to anyone who could make Hugh Loebner desist from holding the contest), but it has also had support from serious intellectuals. Ned Block appears to have been one of this year's judges; Daniel Dennett chaired the panel during some of the early years (but eventually withdrew when he could not get agreement to his plans, which would have seen a number of more 'serious' AI challenges introduced as preliminaries to the main event). It's certainly an entertaining event - sometimes the transcripts of the conversations have a demented but irresistible humour about them - but I wonder how Alan Turing would have felt about it. Nowadays the contest rather underlines the failure of Turing's prediction - that we should have conversational computers by the end of the twentieth century. Personally, I think the other two points which come across most strongly from the event are the continuing weakness of the chatbots and the unserviceable qualities of the Turing test itself.


I think that's a bit hard. There's a vicious circle here, isn't there? The AI panjandrums don't participate because they fear the bots will perform badly and disgrace them; the bots continue to perform badly because they don't get the input of the AI panjandrums. What are Minsky and the rest so haughty about? Designing chatterbots is a perfectly legitimate way of researching several interesting problems; parsing (grammatical analysis) most obviously, but also human interaction in general; the rules of topic maintenance or change; emotional tone; and perhaps even the pragmatics of conversation. It's not hard to imagine bots that made a real contribution to these fields even if they didn't seem particularly human.


I don't think the 'panjandrums' are mainly worried about being disgraced by the poor performance of the bots. The real problem is that there's a kind of dishonesty about the whole enterprise. Nobody is able to produce anything like a real simulation of a human brain in conversation, so the contest implicitly invites people to rig up a program which *appears to be conscious even though it unarguably isn't*. I don't think that's quite what Turing had in mind (although I don't think he ever actually says a machine would have to be conscious to pass his test). Personally, I also think it would be an extremely dangerous thing to do if it were possible, blurring the distinction between people and things in a way which might have disastrous moral consequences.

Luckily, of course, it isn't possible. It's quite clear if you read some of the transcripts that the bots have two big problems. The first is that they cannot follow the thread of a conversation, or even interpret the meaning of individual sentences correctly much of the time. This is because they only really deal with grammar; to decipher human communications you have to be able to deal with intentionality, with the actual meaning, and none of the designers have found a way of doing that. Second, they just don't know enough. You can stall and evade questions to some extent, but a human interogator soon discovers that the bots completely lack whole areas of information that even the most ignorant human being would take for granted - *is* my toe bigger than a 747? I think it's clear that you can't have human-level conversation without human-level consciousness - or something close to it.


Well, you know, one valuable point the Loebner makes is that the Turing test really is a hard one, and maybe it's a bit too hard. I'm sure the reason Turing proposed it because the power and flexiblity of language mean that a conversation can probe the memory, intelligence, and other mental qualities of an experimental subject in a far more searching and challenging way than any other experimental set-up. Maybe we need some scaled-down challenges: I believe that in some past years, the Loebner conversations were restricted to particular topics, for example. Myself, I wonder if it would be more productive to have sessions in which the interlocutors were not actively trying to catch the computer out. Conversations in which the judge merely throws in strings of nonsense to see whether they are recognised as such don't seem to achieve very much.


Yes, but you see, that's the whole problem with the Turing test principle. If you find a group of people who want to believe the computer is talking sensibly, and they make enough allowances for it, you can easily get a positive result. A program which just bats back people's own input in the form of questions, like Joseph Weizenbaum's famous Eliza, is quite capable of fooling some people. On the other hand, if you have a skilled forensic examination, it's always going to be possible to find inconsistencies in the conversation of any human-like entity which hasn't actually lived a genuine human life. So what's the point?


The point is that the prospect of talking to a computer, or a robot, is what actually fires people's imagination. It's like the space program; there's tons of interesting science to be done with satellites and probes, but what actually gets people's interest is *going to Mars*. Like it or not, the programme needs that excitement; and the project is a perfectly legitimate and valuable one, too. It's the same with the Loebner: OK, there are lots of less glamorous projects in AI which in some ways probably deserve attention more than the creation of chatterbots. But if it weren't for the ultimate prospect of a friendly chat with your metal pal, how much interest and support would the artificial modelling of mental processes attract? Heavens - the whole field might be left to neurologists!

Panpsychism

15 October 2005

Am I unfair to panpsychism? Joseph McCard tells me he thinks it is not a 'strange idea' at all.

Subjectively at least, it seems strange to me; in fact, the idea that everything has within it at least a spark of consciousness strikes me as faintly nightmarish. So when I walk down the road, I'm treading on countless minds; pushing my way through gassy clouds of, well, conscious entities? Not an agreeable thought. Rationally, more to the point, I still think panpsychism looks like a dead end. The way I see it, the problem before us is to account for the phenomenon of consciousness, which seems an inexplicable anomaly in the mechanistic world described by physics. Panpsychists make the problem easier by taking a different angle - maybe it isn't an anomaly after all, but a ubiquitous property of everything; it just looks like a special property of human beings from a human being's perspective. This is a neat piece of footwork, but in my view the problem it leaves behind - how to account for the difference between our form of consciousness and the form possessed by a lump of iron, say - is just as bad as the one we started with; in addition, we incur the huge metaphysical commitment of all that consciousness out there which is lying around apparently doing nothing.

I think this, or something like it, is still the orthodox view in philosophical circles. Panpsychism is still often regarded as absurd - so much at odds with common sense that it discredits any theory with which it is associated. The 1987 *Oxford Companion to the Mind* brusquely equates panpsychism with animism, a 'belief common among primitive peoples...' This is inaccurate as well as inadequate: besides ignoring the more sophisticated forms of panpsychism, it overlooks the fact that Shinto, for example, is animistic but not panpsychist: certain salient creatures and objects are considered to enjoy human-style personality, but it's not believed that *everything* is animate.

But panpsychism also has a number of friends and moreover does seem to have enjoyed a significant growth in popularity over recent years. Joseph says there were no Google hits for 'panpsychism' in 1999, about 700 in 2002 and 40,000 now, and it certainly seems to me that there are far more serious discussions of the subject now than there were ten years ago. New reasons for looking at panpsychism again have emerged; if consciousness arises naturally out of information, and everything is covered in information, then it seem logical to expect at least the beginnings of consciousness everywhere. Many people who would not quite commit themselves to panpsychism, like David Chalmers, nevertheless find it an interesting hypothesis.

Paraphrasing a bit, Joseph challenges three of the things I said in the earlier piece. First, he does not consider the empirical unprovability of panpsychism is necessarily a problem; lots of things appear to be unprovable, and a panpsychist's knowledge of consciousness is not based on material evidence, anyway, but on subjective experience. I can see that one's knowledge of one's own consciousness is like that, but I'm not sure how you can subjectively know about the consciousness of other objects, unless by some mystic experience. The claim to have such an experience would not in itself be illegitimate, but it would be optimistic to think that your own private experiences could be convincing to those who haven't had similar ones.

Second, Joseph claims Occam's Razor is on his side, because panpsychism allows you to believe the world consists only of consciousness, a very large simplification. I don't know about that: it sounds more like idealism than panpsychism, as if we've moved on from the view that everything has a mind to the distinct view that everything *is* a mind. I believe it's not uncommon for panpsychist theories to have a tinge of idealism - but to me it seems another kettle of fish altogether, and not therefore an advantage of panpsychism per se.

Third, Joseph addresses the point about what entities qualify for consciousness: is my foot separately conscious? If a brick is conscious, is half of it conscious separately at the same time? I don't mean to present this as a knock-down argument against panpsychism: but it's another problem which I think panpsychists have to deal with. There are various ways this might be done - by distinguishing somehow between real individuals and composites, for example, or by setting up some hierarchical arrangement. Every option has its attractions and its problems, but whichever we take, we incur a further explanatory burden.

So I'm not convinced: but I think I would concede that I haven't done the range and subtlety of panpsychist theories anything like full justice. David Skrbina, earlier this year, published a substantial survey of the history and influence of panpsychistic thought (*'Panpsychism in the West'*): I'm afraid this is among the many interesting books I haven't yet read, but his article in the JCS (Volume 10, number 3) covers similar ground more briefly.

The first point to acknowledge, perhaps, is that panpsychism does come in many forms. At one end of the spectrum there are indeed more or less animistic views which attribute fully sentient, human-style souls to trees or rivers and so on; somewhere off to one side there are theories which regard the sentience of all individual things as offshoots from a pantheistic universal mind; and more popular in recent times, there are panexperientialist ideas which confer on material objects only a basic capacity for phenomenal experience. This range of views illustrates a point made by Skrbina, namely that panpsychism is not, in itself, a theory of mind, but a view which can coexist with many different theories. It tells us that minds are everywhere, but we can hold any number of different views about their nature.

Skrbina's comprehensive historical survey certainly succeeds in demonstrating the existence of a large number of adherents and sympathisers for panpsychist ideas, though there doesn't seem to be any sense of a panpsychist tradition; it's more a matter of a sequence of individual enthusiasts. In some cases, he finds interesting but (to me at any rate) obscure panpsychists like Francesco Patrizi, who apparently has an ontology with nine different levels; in others, he draws attention to the panpsychist views of people (Cardano, for example, and Fechner) who are slightly more familiar, but whose claims to fame actually stem from largely unrelated intellectual achievements. But he also provides a novel perspective on a number of major figures. I don't know why, but I don't think I'd ever really thought of Leibniz in a panpsychic context. But if the world ultimately consists of monads, and monads are souls, then it surely follows that the world consists of souls. I tend to think of Leibnizian monads as generally more akin to points of view than to full-blown human souls. This suggests another possible property of the spectrum of panpsychist views; that the ontological commitment involved gradually diminishes as we progress across it. To believe that a mountain has human-style thoughts and intentions requires some quite hefty additional beliefs about the world; to believe that everything has the basic qualities required for a dim glow of experience is less of a stretch. If you downgrade the commitment enough, at some point you end up with a theory which is undoubtedly true, but has ceased to be very interesting (or very panpsychistic for that matter): you merely believe that everything has the capacity to be part of a mind.

Could it be that a gradual drift in this direction will eventually eliminate genuinely panpsychist ideas? Not, I think, any time soon; there's a lot more to be said on the subject yet.

Almost a Virgin

31 October 2005

David Chalmers recently commented that Jaegwon Kim's new book 'Physicalism, or something near enough' showed how old-fashioned materialism was slipping in the popularity stakes: not suffering wholesale rejection exactly, but no longer enjoying the status of a near-unchallenged orthodoxy.

Kim himself says that the book offers no startlingly new views, only a clearer and better argued case. But the interest of some readers has at least been distinctly quickened by the main conclusion; that physicalism is nearly, but not quite, the whole truth. How can that be? At first sight, claiming that one is nearly a physicalist is about as plausible as claiming that one is nearly a virgin: in such cases even the most progressive logicians would generally exclude the middle. How can you have more than one fundamental substance without having two?

On reflection, however, I can see some attractions in that way of thinking. After all, physicists and mathematicians sometimes talk about spatial dimensions in fractional terms. If one can coherently have a space which is not 3- or 4-, but 4½- dimensional, why not fractional substances? What about a graduated ontological scale? We might say that anything over 1.5 counts as dualist, while anything less passes as monism; substance dualists would be up in the 1.8 to 1.9 range, while property dualists generally scored around 1.6 to 1.7. Only the most obdurate materialists would manage a 1.0, while some bold souls - Penrose, Popper? - would score values in excess of 2.0. Scores below 1.0 would surely be unattainable, unless perhaps reserved for the softer kinds of idealism ('life is but a dream').

In fact, joking aside, I really think that the scope for vagueness provided by such a scale might depict the discussion in realistic terms. It seems to me that if the fundamental argument over dualism is interpreted too strictly, there is a straight choice between monism and epiphenomenalism. If your second fundamental stuff is really detached from the first, it must be an irrelevant epiphenomenon which you might as well forget about; if, instead, it has at least some causal connections with the first, you might as well exercise ontological economy and regard both stuffs as part of a single comprehensive world (after all, believing in energy as well as matter doesn't make you a dualist; so why should believing in suitably causal spirits?). A general application of the latter approach would, of course, have the strange effect of reclassifying nearly everyone, including Descartes, as monists - whether they knew it or not.

In practice a dualist is not generally someone who believes in the utter, unbridgeable bifurcation of the universe, but someone who believes that a twofold division of some kind is especially salient: more important than the distinction between energy and matter, but not so important that the world falls apart down the middle. The precise degree of salience might indeed be a matter for a graduated scale. Perhaps there's actually less ontology involved than meets the eye: perhaps we should be better off, in most cases, if we saw the choice of monism or dualism as more a matter of expositional strategy than fundamental principle. One reason we don't, I suspect, is that people have a strong desire to exclude ghosts and Christian spirits from their theories, and dualism (which need not actually involve souls or spirits at all) gets used as a

proxy. Perhaps that's also why Descartes keeps on getting such a bad press: not really for dualism per se, but for a dualism which explicitly made room for Jehovah.

So then, maybe in principle it's alright to be nearly a physicalist - so long as the theory itself is OK. What is the nature of Kim's dualism, or small lapse from perfect physicalism?

The bulk of the discussion is taken up with establishing Kim's basic physicalism, with the 'lapse' discussed mainly at the end of the book. If we want mental events to have real causal power (and we surely do), Kim argues that they must be reducible in principle to the physical level. He carefully distinguishes this view from eliminative reductionism: phlogiston, he says, was eliminated, and thereby banished from science, while heat and temperature were reduced, and remain useful concepts. Moreover, after a careful survey of the options it turns out that the only kind of reduction available is a functional one.

This is all fairly persuasive, though not so seductive that it is likely to make many converts, and it leads us to a pretty familiar position: it turns out that many mental events are susceptible to functionalist reduction and can therefore be brought within the pale of physicalism. We may not be able to spell out the account in full, but we have a pretty good idea of its broad outline. The problem, as ever, is qualia. Qualia are not susceptible to functional reduction (the possibility of inverted qualia is sufficient to establish this: Kim rather loftily remarks that we don't need any of that zombie stuff). The distinctive feature of Kim's analysis is that he nevertheless wants to take another slice off the unsolved problem; claim the border regions of the qualia country for physicalism. This he does by asserting that while qualia in themselves cannot be characterised in functional terms, differences between qualia can. So, it would not matter in practical terms if the real experience of red and green were switched: we should still be able to interpret traffic lights correctly. But if there were no difference between the real experiences, we should be in trouble. Thus qualia do have a real causal role, though in their essence they remain ineffable.

It's this move which allows Kim to claim he's still really a physicalist; even qualia have been naturalised and given a role; the fact that they still have an inscrutable phenomenal aspect isn't a cause for concern. The tone of Kim's conclusion makes it clear that he regards this as an acceptable final position; he isn't in despair over qualia and he isn't looking for more to say about them, either.

This seems a little strange, even on Kim's own terms. An inexplicable feature of the world just demands philosophical attention: you expect a philosopher to respond to a hint of the unexplained in the way a mother responds to a faint sound of crying. In Kim's eyes, the problem has been substantially reduced; but half a problem is still a problem. I do have some sympathy with idea that people worry too much about qualia, but that's because I think there's an element of confusion about the whole thing: a kind of clerkish surprise that accounts or explanations of experience don't bring with them the experience itself. On Kim's account, a significant portion of the old problem is still there: he just somehow manages not to worry about it.

I'm not convinced that the argument about differences really works, anyway. In the first place, is it really true that qualia play a causal role in our response to traffic lights? It's certainly possible to detect the difference between red and green, and respond accordingly, without any trace of qualia: quite simple machines could do so, and it rather looks as if the brain uses some broadly similar mechanisms. I suppose we could argue that qualia over-determine or somehow help out with the response, but it seems easier to conclude that these functionalisable qualities which Kim is talking about are not really qualia at all.

Secondly, doesn't our ability to spot differences between qualia depend causally on our ability to perceive them in the first place? It seems to me that we see the difference between red and green qualia only because we actually see green and red. But if that is so, and if our perception of qualia differences has a causal role in determining our behaviour, then qualia themselves also have a causal role, albeit one step removed.

I suspect there is something in what David Chalmers says about a noticeable swing away from monist materialism; but I also suspect that the reasons are more to do with despair over the current impasse than any new and stronger case for dualism. You could see Kim's theory as another desperate effort to find a novel yet plausible way forward - an effort which nearly succeeds. Unfortunately, in this case, nearly really is the same as not at all.

Spin Doctors

10 November 2005

Huping Hu and Maoxin Wu have recently produced a new paper elaborating their views about the fundamental role played by quantum entanglement in consciousness, among other things.

Entanglement is certainly a strange affair. As I understand it, when two particles become entangled, the state of one can't be measured without implications for the state of the second: the strange thing is that this state of affairs can persist even when the two particles become spatially distant. This leads to the dreaded 'spooky action at a distance', where measuring one particle appears to have an instant effect on another which is far away even though there was no time for any kind of signal to have travelled between them. No less a person than Einstein found this unsatisfactory, and at first sight it certainly seems to contradict common sense and the upper limit imposed by the speed of light.

The correct interpretation of this phenomenon, and of the rest of quantum physics is notoriously controversial, of course. Hu and Wu appear to believe that the apparent action at a distance requires (or at any rate, uses) a special realm, pre-spacetime. Pre-spacetime, rather like theological Eternity, has no duration, nor extension in itself, but remains adjacent in some sense to every moment of time and point of space. It is this strange kind of universal proximity which allows quantum effects to act into pre-spacetime and then from there influence events elsewhere (at least, that's how it works if I've understood correctly).

My opinion on quantum physics is worth nothing, but I doubt whether entanglement is really as spooky as it seems. Consider this analogy. We deal out the cards for a hand of whist. Three sets of cards remain on Earth, while the fourth player takes his on a spaceship and heads off to the region around Betelgeuse. When he is deemed to have arrived, the cards are taken out again (presumably by the descendants of the original players), and following the rules for this particular variant of the game, the first player on Earth nominates the suit which is to be trumps. At that moment, the hand of cards out by Betelgeuse becomes a winning one (if played correctly), or perhaps a losing one. How can it be that the decision about trumps on Earth instantly affected the status of the cards many light-years away?

Alright, perhaps it isn't quite like that, but I'm inclined to doubt the need for the additional ontological commitment implied by pre-spacetime, and on those grounds I should prefer some other interpretation. But perhaps if the theory provides a good angle on consciousness it is worth having after all - so what are Hu and Wu actually putting forward?

They put spin at the centre of their theory, and suggest that "the nuclear spins inside neural membranes and proteins form various entangled quantum states and, in turn, the collective dynamics of the said entangled quantum states produces consciousness through contextual, irreversible and non-computable means and influences the classical neural activities through spin chemistry".

Once again, pre-spacetime plays a crucial role. It is there that consciousness actually resides, enjoying two-way communication with the brain via entanglement. I suppose this unification of

the mind in the non-spatial realm of pre-spacetime could be offered as a solution to the binding problem, though as I think I have suggested before, merely proposing a mechanism capable of linking everything more or less indiscriminately doesn't really provide an answer. Placing the mind in a kind of eternal world like this does have some resonance with traditional beliefs, but it must also involve Hu and Wu in the kind of philosophical difficulties which have always arisen in connection with an eternal God's ability to intervene in a temporal world.

But even if we ignore the philosophical problems, there also seems to be a technical objection: it hasn't been established that information can be transmitted by entanglement alone, and there is fairly good reason to think that it is impossible. Hu and Wu acknowledge this but suggest that an entangled entity can sense and use the entanglement directly. Can the 'entity' they have in mind be a mere particle? Unless we are to invoke some variety of panexperientialism, I don't really see how that would work. They say that "spin is the mind-pixel", which suggests that they are primarily interested in consciousness as awareness. The idea that we have a picture made up of pixels on a screen in our minds, however, is rather strongly reminiscent of the dubious 'Cartesian Theatre' which Daniel Dennett has expended so many words denouncing.

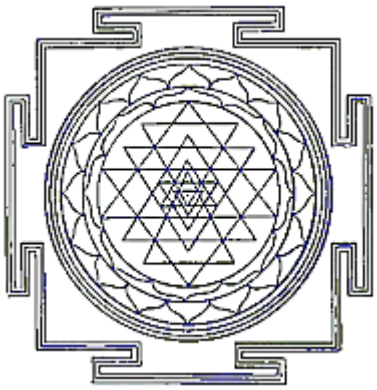
I mentioned that the theory could perhaps be presented as a contribution to solving the frame problem, but that isn't a claim made by Hu and Wu, and in fact it doesn't seem altogether clear what motivates their theory - why they think the quantum mechanism provides a better explanation than classical systems could do. They describe the operation of the mind as non-computational, but I don't believe that follows automatically from the invocation of quantum physics.

A positively alarming feature of the paper is its willingness to consider some other, rather wilder, potential implications of entanglement. Because of its supposed property of action at a distance, it might have a role in explaining parapsychological phenomena (if there are any); it might help explain the efficacy of homeopathy (if it is effective) - specifically by providing a basis for the homeopathic idea that the properties of water are affected by a 'memory' of substances it has previously encountered. Personally, I think the theory that the properties of water depend on its past history must be among the most extensively refuted hypotheses ever: every chemical experiment involving water implicitly confirms that its properties are constant.

Could it be that these stranger claims (although 'claims' is too strong: Hu and Wu offer all their ideas only with the most graceful diffidence) offer a clue to the motivation for the theory? Could it be, in fact, that the very spookiness which Einstein found intolerable is the feature that Hu and Wu find attractive? Perhaps they hope that quantum theory will be able to take the place of the magic and even the spirituality which featured so strongly in pre-scientific accounts of the world. If so, I think they are doomed to disappointment: it may seem strange and difficult to us (it certainly does to me), but quantum physics is as rigorously scientific as Isaac Newton.

Vedic Consciousness

20 November 2005



Prashant Singh proposes (pdf) an hierarchical theory of consciousness. He suggests that Western investigations of consciousness may have been too narrowly constrained within a reductionist framework and might benefit from a transfusion of ideas from other sources. The main ideas in the present paper drawn from Vedic thought seem to be first, that consciousness is non-material, and second, that it is indeed the topmost layer in an hierarchical structure. Neither of these in themselves seem to me particularly radical additions to our stock of ideas, though they constitute a potentially interesting position. However, it is also true that the general perspective of the paper, which sees human beings as possessing a special kind of original causality, is rather different from that of most modern discussions – in fact it reminded me to some degree of typical ancient Greek views of the soul (which might well be closer to Vedic ideas than the contemporary view).

It may be that further and more specifically Vedic ideas would appear in the explanation of how the non-material consciousness and the material brain interact; but that crucial part of the theory is deferred for the moment.

Prashant distinguishes three levels in his structure, or four if you count the simple physical body. Above the body is the mind, still a physical and more or less computational affair; intelligence is a governing level above that, though still essentially a function of material processes; and finally non-material consciousness.

In Prashant's view, a crucial difference between us and computers (at least, those that exist at the moment) is that we have kind of self-programming ability which they lack. A computer can stolidly record data: it can deliver outputs for inputs as specified by its program, but it can't alter its own program while it's running, and as a result there's a kind of unpredictability which is beyond it. Is this true? In a narrow sense I think it probably is true of most real programs; but I'm not sure it captures a genuinely significant difference between two kinds of process. Prashant thinks the results of a straightforward program can be directly predicted, whereas the results of a self-modifying program can only be predicted by actually running the program. I'm not sure: I don't really see how you can foresee the output of any program, even a simple look-up one, except by doing something functionally equivalent to running it – even if you only 'run' it in your imagination. But I think the formal argument is really only there to buttress an intuitive conviction that our mental processes are creative or original in a way which narrow computation does not seem to capture.

Perhaps the most interesting argument is one about the origin of knowledge and intelligence. Prashant draws an analogy with the data storage and processor of a computer (Debatably, I think: surely the intelligence of a computer is at least partly attributable to the program? But the point isn't really material to the argument.). Given a particular state of knowledge and intelligence, he asks, can we work out what the previous states might have been? Well, at least we can say this much: that a state involving some material level of knowledge and intelligence cannot have developed directly from a state in which there was none of either. It follows that

mentality only comes from mentality, and that any mental powers possessed by machines are properly attributed to their designers or creators. On the face of it, this argument should make the birth of human beings with their own mental life impossible: but the regress is broken, in Prashant's view, by the intervention of non-material consciousness. If it weren't for this intervention, we should be involved in a disastrous infinite regress.

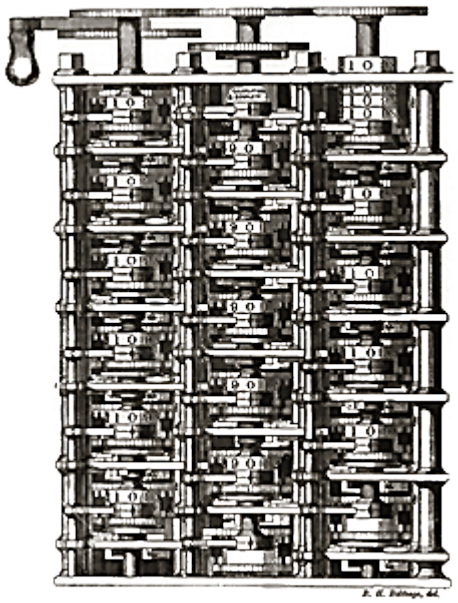
This argument reminds me of two similar ones, one bad and one good. First, it looks a lot like the various arguments for God's existence which would have him as the first step in an otherwise infinite chain: he is the first cause which explains all other causes, or the necessary being which explains the existence of all the contingent entities, including ourselves, with which the world is filled. The problem here is that God also seems to require a cause, and no adequate argument for his logical or metaphysical necessity is available. Simply declaring him the first link in the chain doesn't provide the explanation we need: and similarly, declaring that our knowledge and intelligence spring from a non-material world doesn't deal with the original problem. So the physical knowledge came from non-physical knowledge: and where did the non-physical knowledge come from?

The good argument which Prashant's view reminded me of is one about moral responsibility. If we wanted to create a conscious robot, we should somehow have to stand aside from its design; otherwise, if we specified how it was to work, the moral responsibility for its wholly foreseeable actions would remain ours, and it would fall short of human consciousness at least by failing to have the common human capacity of responsibility. If robots made other robots and those robots made others indefinitely, the moral responsibility would travel back indefinitely; but in this case the regress is interrupted by human beings who were born, not made, and have original responsibility just because they arose out of a designless process of evolution. The solution in this case is Darwin rather than God, and evolution rather than a non-material world.

I'm not convinced by the main arguments, therefore, but I do have some intuitive sympathy with the general idea of an enquiry into what look like the special causal powers of human beings. Prashant argues that the Turing test is useless because no mere measure of behaviour can identify consciousness: consciousness implies an ability to behave, or refrain from behaving, in ways which are not pre-determined by our inputs. I think there's some truth in that.

A Consciousness Engine?

2 December 2005



I can't help wondering what Charles Babbage would have to say on the subject of consciousness if he could somehow be brought back to life. Such a formidable thinker, with the unusual combination of creativity and rigour needed to come up with ;the basic ideas of computing all on his own: he would surely have interesting things to say about the prospects of Artificial Intelligence? Indeed, someone who could conceive first the calculating Difference Engine, and then the computing Analytical Engine, might surely have had ideas about how the final steps to a thinking Consciousness Engine might have gone?

Babbage has often been unfairly portrayed, even in his own lifetime as hopelessly impractical: seduced by a new brilliant idea before he could bother putting it into practice. It is true, however, that the construction of the Difference Engine cost huge sums and was never completed. One reason, I suspect, is the sheer size of the thing. You would expect a calculating machine to resemble a clock, or some similarly delicate mechanism, but the Difference Engine was constructed with massive brass gears and hefty shafts as though it were going to be part of the lifting gear for Tower Bridge. This mighty scale is certainly not essential to the calculating function - both Babbage himself and a Swedish enthusiast who had read about the Difference Engine made smaller, less ambitious models which worked perfectly well - but it must have drastically increased both the cost and the engineering problems. Of course, we have to remember that the Difference Engine was intended to crank out mathematical tables on a special printer, so it actually was meant to be a kind of factory machine, and therefore needed to be pretty robust. In a sense this was an essential feature: the whole point of the machine was to eradicate human error, and having the results transcribed manually would have allowed mistakes to creep back in.

Another reason for the failure of the attempt, to judge by the interesting account given by Doron Swade in *The Cogwheel Brain*, was that Babbage's chief engineer Joseph Clement was not the ideal man to deliver a project: a genuinely brilliant engineer, but a rather argumentative perfectionist, disinclined to let costs stand in the way of perfect results. Doron Swade, of course, was responsible for the late vindication of Babbage's design, and implicitly of Victorian technology, which was achieved when the Science Museum in London finally produced a full-scale Difference Engine which functions perfectly

It's really the Analytical Engine which is of interest to us, though. Babbage borrowed from Jacquard the idea of using punched cards as a way of programming complex behaviour and endowed the new machine with the power to do conditional branching and looping; he also made the imaginative leap from seeing it as a calculating machine to the more general conception of a formal symbol manipulator. If it had been built, the Analytical Engine would have qualified as the first true computer, an astonishing achievement given the relatively simple calculating machines which were its only real predecessors (In passing, it's interesting to note that Leibniz was responsible for one of these. Given his concern with logical calculation - and

binary numbers, as it happens - I wonder whether he influenced Babbage's ideas in this respect. Babbage was certainly an active campaigner for the adoption of Leibnizian notation for calculus, other English mathematicians of the day clinging obstinately to the Newtonian approach.)

Unfortunately, there aren't many sources for Babbage's view on the wider implications of computation, and one of the main ones wasn't written by him. This is Menabrea's description of the Analytical Engine, and particularly its translation, with additional notes, by Ada, Countess of Lovelace.

Ada, the daughter of Lord Byron, worked closely with Babbage, and views about her differ sharply. She has been celebrated as a pioneering female genius and recognised as the first computer programmer on the strength of some routines for the Analytical Engine which are described in her notes. A programming language has been named after her, and she has become a significant figure in feminist historical analysis (notably as portrayed by Sadie Plant in *Zeros and Ones*). On the other hand, she has also been described as 'mad as a hatter', 'the most over-rated figure in the history of computing', her mathematical skills likened to those of a novice student, and the famous Notes attributed largely to Babbage himself, with Ada no more than an aristocratic, high-profile secretary. It's true that she never produced any other interesting work on computation or mathematics, and while it's always difficult for third parties to interpret personal letters correctly, Ada's undoubtedly suggest someone whose ego was vastly out of control.

Be that as it may, the Notes themselves are lucid enough and it seems fair to assume that Babbage at least agreed with what they say. The key passage from our point of view occurs in Note G.

> *"It is desirable to guard against the possibility of exaggerated ideas that might arise as to the powers of the Analytical Engine. In considering any new subject, there is frequently a tendency, first, to overrate what we find to be already interesting or remarkable; and, secondly, by a sort of natural reaction, to undervalue the true state of the case, when we do discover that our notions have surpassed those that were really tenable.*
>
> *The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths. Its province is to assist us in making available what we are already acquainted with."*

This looks like a disavowal of any power of thought on the machine's part, and by implication on the part of any computer, but perhaps it isn't. In the first place, Babbage was exasperated by some of the questions he was asked about the Difference Engine - could it get the right answer even if it were given the wrong data? - and it might be that Ada was ruling out similar semi-magical powers on the part of the Analytical Engine. In the second place, Alan Turing thought the Note did not necessarily mean the machine couldn't think. The last part of the passage quoted appears in his celebrated paper *Computing Machinery and Intelligence* as 'Lady Lovelace's Objection'. Perhaps in a sense, he says, the Analytical Engine would have had the power of thought: it was, in his sense, a universal digital computer, so if any computer could think then, subject to constraints of speed and capacity, the Analytical Engine could. He responds to the point about the machine's lack of originality by saying that machines can in fact

often produce unexpected and surprising results, even for those who built and programmed them.

Another possible source from which we might glean Babbage's view is the Ninth Bridgewater Treatise, his riposte to a series of earlier essays, in which Babbage tried to defend science as compatible with enlightened orthodox religion. Here there's no doubt that we are reading Babbage's own views; the problem is that he isn't primarily writing about the capacities of machines.

However, there are some interesting points here. Babbage argues that miracles need not require one-off interventions from God: in creating the world he could have set it up in such a way (it's hard not to say 'programmed it') that the apparently exceptional behaviour is really part of the laws of nature, which are just a bit more complex than meets the eye. He points out that he could set up a calculating machine to crank out the numbers in a mathematical series which looks perfectly regular, but at the tenth, or hundredth, or one million four thousand and forty-second place, produce a single anomalous figure, or indeed switch to calculating a different series. For that matter, it might be the same with animals: it might be programmed into a given species to live only a certain number of generations and then morph into another form or simply die out (most modern readers will feel this idea has been superseded by the work of Babbage's friend Darwin).

These passages show how a kind of computationalism had come to influence Babbage's general view of the world; it seems natural to him to think that the laws of nature and biological heredity simply express the working out of algorithms set up by God. Demoting the laws of nature to a set of current rules whose operation might in principle change in any way at any moment is actually a slightly unsettling piece of metaphysics.

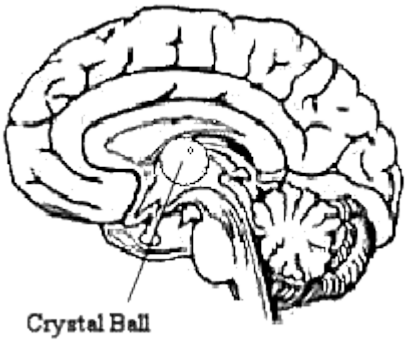
At the end of the Treatise, he mentions the subject of free will: he declines to discuss such a thorny philosophical issue directly but offers one thought which might be found helpful. Again, he uses the example of setting up a calculating machine: it's possible, he says, to have a certain result occur only in certain conditions. For example, ' *when calculating a table of squares, it may be made to change into a table of cubes, the first time the square number ends in the figures 269696; an event which only occurs at the 99736th calculation.'* But the person setting up the machine need not know that 269696 will come up on the 99736th calculation: in fact, they needn't know that it ever comes up in the normal course of the calculation at all.

In other words, prefiguring what Turing said, you can still be surprised by the outputs of a machine you have programmed yourself. Turing suggested that being surprising was the nearest we might be able to get to real originality or freedom, and a similar thought appears to be in Babbage's mind. There's a difficulty about applying this argument to God and free will because of God's omniscience: he can never be in any doubt about when 269696 is going to come up, or when human beings are going to do something surprising; but it does suggest that a machine could be free and even original in some sense, in spite of being built and programmed. In fact, since Babbage presents the point as relevant to human free will, there is an implication that Babbage thought human beings were at some level computational entities, albeit ones programmed by God.

That's quite an inference from what is really no more than a hint, of course: I feel fairly sure Babbage was aware of all these implications, but that doesn't mean he accepted them, even privately. Alas, we shall never know for sure.

Feelings First

11 December 2005



The idea that quantum physics might have something to do with consciousness crops up in a number of theories: perhaps most prominently in the version put forward by Penrose and Hameroff, but quite regularly elsewhere - the views of Huping Hu and Maoxin Wu, for example, have been discussed here recently. I sometimes find the motivation of this line of reasoning a little unclear: why exactly do we need to invoke quantum physics, anyway? Sometimes it seems that the quantum theorists are merely hoping that there will turn out to be one solution to two otherwise unconnected mysteries: not actually a very good bet.

Earlier this year, Maurits van den Noort, from the Department of Cognitive Neuroscience at the University of Bergen, came up with yet more reflections on the relevance of quantum effects, but this time there is a difference, in that he puts forward two studies which seem to offer empirical evidence that something unexpected and possibly quantumish is going on in the unconscious emotional reactions of human beings, as revealed by scans and EEG readings. A book is apparently in the pipeline.

Van den Noort gives an admirably clear summary of just how quantum physics might be thought relevant to consciousness:

> *First, quantum coherence (e.g. Bose-Einstein condensation) is a possible physical basis for 'binding' or unity of consciousness (Marshall, 1989). Second, non-local entanglements (e.g. 'Einstein-Podolsky-Rosen correlations') serve as a potential basis for associative memory and non-local emotional interpersonal connection. Third, quantum superposition of information provides a basis for preconscious and subconscious processes, dreams and altered states. Finally, quantum state reduction (quantum computation) serves as a possible physical mechanism for the transition from preconscious processes to consciousness (Penrose, 1989; 1994).*

He says that quantum state reductions may send quantum information "back in time", and it is the scope for non-linear information processing which particularly interests him. Acknowledging that a smooth chronological sequence is one of the more salient properties of subjective experience, he points instead to unconscious information processing, and especially to the generation of emotional responses.

The two experiments quoted by Van den Noort involved showing subjects a series of images: some neutral, some emotionally stimulating. (In fact the stimulating images were either erotic or violent, which seems to short-change the emotional life of human beings somewhat, though no doubt it helped to guarantee a clear reaction). Both experiments, whether based on scanner or EEG, produced the same extraordinary result: it was possible to detect an emotional reaction to emotional stimuli just before they actually occurred, even though subjects had no way of knowing whether the next image in the random series was going to be emotionally charged or neutral.

Van den Noort suggests that a likely explanation is that the unconscious mental processes at work here are powered by non-linear quantum information processing. He sees unconscious

information processing as a kind of fast but rough system for generating emotional reactions which colour our behaviour, although they remain subject to the slower and more thorough examination of events which takes place on a conscious level.

These results are so remarkable (I'm not sure I can get my head round the implications for Libet's similarly counter-intuitive findings at all) that it is tempting to assume that something must have been wrong with the experimental set-up: perhaps the images weren't truly randomised, or the subjects were somehow able to pick up subtle clues of which the experimenters remained unaware. I don't know of any flaws like this, but I am inclined to think something must be wrong somewhere.

First, the ability to respond to future events, even in the shortest term, is such a useful capacity it seems strange our nervous system doesn't make better use of it. Our system is certainly set up to produce fast responses in certain cases: when we touch a hot surface, for example, the nerve signal does not even have to spend the small amount of extra time needed to go all the way to the brain: instead, a short loop triggers a hard-wired evasive response. If the brain could register emotional events before they happened (and burning your hand, or even nearly burning it, surely does generate some emotion) it ought to be able to set off the reflex evasive action before you actually make contact. I realise that evolution does not necessarily deliver every possible good idea: but this one is so simple and so valuable it seems surprising, given Van den Noort's findings, that pre-emptive reflexes don't exist. Second, this kind of experiment, monitoring reaction times to a sequence of images, sounds like a pretty typical piece of psychological research: it seems a little odd that some other researcher hasn't come across this chronological curiosity before.

Still, there is a chance - just a chance - that this represents a breakthrough of some kind.

Database

14 January 2006

Our conceptions of how the brain works have always been conditioned by the available technology. Descartes spoke of hydraulic automata; Leibniz spoke of mills (and Babbage named one of the core components of the Analytical Engine the 'mill'). I'm old enough to remember when a telephone exchange was a popular analogy for the brain - people still, in a rather similar vein, suggest hopefully that the Internet might develop some mental properties when it attains a sufficient level of complexity. But for many years now, the favoured analogies have been derived from computers, whether the parallel is drawn with hardware, software, or both. Clearly the question of how adequate these computational analogies really are is still a live one. There are plenty of evident differences between brains and computers, but even so, the computer analogy may still be the best available. As Fodor put it, remotely plausible theories are better than no theories at all.

Ken Brown has offered a new computational analogy in his recent paper - he suggests that mental functions are strongly analogous to those of a relational database. Or rather, some of the higher functions: he draws a distinction between simple neuronal data, the sort of signals which go back and forth along nerves, reporting sensations registered by the skin, for example, and proper, structured data of the kind you might expect to find in a database. He suggests a plausible hierarchy of methods by which our behaviour is controlled, from reflexes through well-learned routines to emotional responses and careful conscious thought - the more time you've got, the more sophisticated the system that gets applied (I'm not sure conscious decisions are *always* slower than emotional responses, but that doesn't affect the general point).

The surprising aspect of the theory is that it incorporates a kind of dualism. The suggestion is that most of the observable activity in our brains does not arise from the data themselves, but from storage and retrieval activity. The neurons are not providing the whole database, only indexes. The data exist in a separate mental realm, and it follows that we shall never be able to pin down direct neural correlates of consciousness. Brown acknowledges that possibility that this other realm could still be one constituted by normal physics, but he really thinks it is most likely to be found in some area provided by an as-yet-undiscovered extension to current physics. He goes to some lengths to remind and assure us of the incompleteness of current physics and the scope for such an extension. This may sound reminiscent of Penrose's reliance on not-yet-discovered quantum theory, and indeed Brown himself quotes Penrose's views.

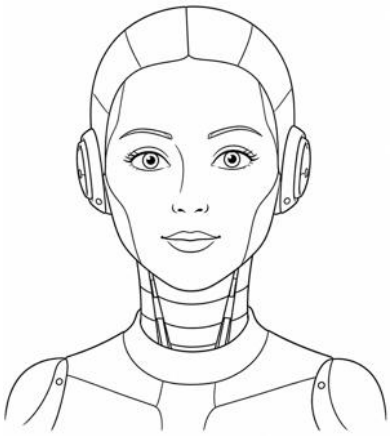
I think Brown succeeds quite well in making his thesis seem natural and plausible, but I have two serious reservations about it (apart from its dualism - though until the desired new physics comes along, the exact extent of the metaphysical commitment is also unclear). One problem is that I doubt whether even the most powerful relational database is quite up to the job he wants from the mind. Brown says we must be dealing with a highly structured system, capable of matching information from a wide range of different sources (and that's why it has to be relational) - but the mind's ability to draw new conclusions, reinterpret whole sets of information, and indeed, come up with entirely new ways of thinking, seems to me to outrun any mere database. Brown argues that all our visualising and reasoning is relational in character -

perhaps there are other ways of thinking, but if so they are unknown to us, and cannot therefore be the way our minds work. This point seems to confuse two different things: the ways of thinking about things which we deploy consciously, and those mysterious unconscious processes which somehow sustain the thinking itself. It also seems a risky point of view, even in terms of mere technology, to suppose there could never be anything better at handling data than the kind of relational database we already have (I believe relational databases themselves were originally considered to be a largely theoretical possibility, never likely to be capable of operating at practical speeds).

The second major problem is that I rather doubt whether the brain really deals in data at all except on a conscious level. Suppose we have a robot which has to reach for an object. One way to set this up would be to gather data about the position of the object, probably expressed as co-ordinates, perform some calculations on the data and compute values which are then fed back into the control mechanism from the robot's arm. But back in 1986, Paul Churchland explained *("Some reductive strategies in cognitive neurobiology", Mind XCV no.379)* how the same task could be accomplished at even greater speed by a "state-space sandwich", a linking of maps which implicitly transformed co-ordinates from the eyes into co-ordinates for the arm, but without doing any explicit algebra or invoking any symbolic representation. It can be argued that the sandwich does encode data, just in an unusual way - but even on that view the encoding is surely radically different from the kind you would need for entry in a database. I think it is likely that all the 'engine-room' work of the brain will turn out to take place through mechanisms broadly like this, at a lower level of abstraction than a computer program would require.

RoboMary

21 January 2006



Bloody but unbowed, Daniel Dennett returns to the attack on qualia in *What RoboMary Knows*, (in *Sweet Dreams* , 2005: also currently accessible in draft form here). I've always thought there was something heroic about Dennett's flat denial of qualia (the ineffable, indescribable subjective qualities of things, the redness of red, etc). Even if you don't share his scepticism, it was surely a point of view which - at least - needed airing. But it seems that keeping the qualophobe flag flying, with few allies, has not been easy: the colleagues who never really address the issue, the annual wave of students who greet the Dennettian position with surprise and disbelief – they take their toll. Dennett therefore, has decided to go back over the ground one more time, spelling out his position more carefully than ever. What's more, he has recruited a new and powerful ally – RoboMary.

As you may have guessed, the discussion is based around the classic thought-experiment (or intuition pump, as Dennett calls it) of Mary the colour scientist. The original argument, briefly is as follows. Mary has never seen colours, but knows everything physical about them. When she sees a rose for the first time, she finally knows what it is like to see red, and therefore, it is claimed, learns something new. Colour qualia, it follows, are something real about the world over and above the purely physical account.

Dennett's response has always been flat denial. If Mary really knows *everything* about the physical aspects of colour (which is harder to imagine than it may seem), then she will be able to tell what it is like, and the rose will come as no surprise whatever. In RoboMary, Dennett recapitulates his blue banana version of the story: Mary is presented with a painted banana but easily spots the trick from her comprehensive knowledge of blue and yellow.

But how could she do that? It's almost as though Dennett believes there could be an algorithm for converting greyscale values to colour ones and automatically colorizing black and white films. Of course, what he really thinks is that Mary knows so much about human physiological reactions to colour that she could distinguish between her own reactions to blue and to red even though she had never experienced those reactions before. But that doesn't seem especially plausible either. Enter RoboMary!

RoboMary is simply the cybernetic equivalent of the original Mary. Although she has been fitted with black and white cameras, she is able to observe the reactions of other robots of the same model as herself and note what values different objects cause to appear in their colour data registers. She then arranges a special routine within herself which uses appropriate values to colorize the input from her cameras, and – Aha! – sees what red is like.

But this is cheating (Dennett acknowledges). Even if we accept that RoboMary is a legitimate analogy for a human being, the comparison breaks down here – we can't voluntarily install subroutines within our optic nerves. So suppose Mary's vision system is protected from tampering. No problem: she simply sets up a virtual model of herself in her own capacious mind and uses it to work out what the state of her brain would be if, when looking at the scene she currently sees, her colour vision were enabled. Then she puts herself into that state. To put it in

more human terms, she imagines what seeing red would be like to a person like her, and that enables her to imagine correctly what seeing red is like.

It's hard not to feel that there is a bit of question-begging going on here. Human beings can't put themselves into a specified experiential state voluntarily. They can't reset the data in their own brains except in certain special cases, and it is a moot point whether human brains have clear 'states' in the way that a computer necessarily does. But really Dennett isn't arguing for a perfect analogy: he just wants to make it seem more plausible that imagining in sufficient detail how an experience would seem can really lead to knowing how it seems.

In a thoughtful response in the JCS, Michael Beaton argues that the real flaw in the argument is that RoboMary has to do something in order to know what red is like. If qualia are just part of the physical world, knowing all the relevant physical facts would in itself be to know the qualia: you wouldn't have to do calculations and research and put special values into data registers – *you'd know already*.

I think Dennett could respond to this by drawing the distinction between what we know and what is actually present in our thoughts. I know what red looks like even when I am in a dreamless sleep: but in order to think about what it's like, I have to go through a process, albeit an instantaneous one of calling it to mind. He could claim, similarly, that RoboMary did implicitly know already what red was like, but that she had to take some special steps in order to think about it, or direct her attention to it.

I think Dennett is wrong for a different reason. Beaton points out that the Dennettian line is quite different from that taken by other qualia deniers such as Paul Churchland and David Lewis. On their view, the problem with the original argument is an equivocation between two senses of 'know'. Knowing what red is like is not the same as knowing what the colours of the spectrum are (you can write the second one down, for example). Knowing what red is like is perhaps a case of 'knowing how' rather than 'knowing that': if you know what red is like, you have the ability to distinguish it from other colours: but you don't possess any more facts than someone who doesn't. Mary, therefore, did not learn anything new in the required sense when she first saw the rose.

I am inclined to think that that is on the right track. I doubt, in any case, whether Dennett's new exposition will be any more effective than his old ones. Arguably the best formulation of his case, both in terms of clarity and persuasiveness, was the two-word one he came up with many years ago: *what qualia?*

Not Like That

30 January 2005

Jerry Fodor's 2001 book *'The Mind Doesn't Work That Way'* makes a cogent and witty deflationary case. In some ways, it's the best summary of the current state of affairs I've read; which means, alas, that it is almost entirely negative. Fodor's constant position is that the Computational Theory of Mind (CTM) is the only remotely plausible theory we have - and remotely plausible theories are better than no theories at all. But although he continues to emphasise that this is a reason for investigating the representational system which CTM implies, he now feels the times, and the bouncy optimism of Steven Pinker and Henry Plotkin in particular, call for a little Eeyoreish accentuation of the negative. Sure, CTM is the best theory we have, but that doesn't mean it's actually much good. Surely no-one ought to think it's the complete explanation of all cognitive processes - least of all the mysteries of *consciousness*! It isn't just computation that has been over-estimated, either - there are also limits to how far you can go with modularism too - though again, it's a view with which Fodor himself is particularly associated.

The starting point for both Fodor and those he parts company with, is the idea that logical deduction probably gets done by the brain in essentially the same way as it is done on paper by a logician or in electronic digits by a computer, namely by the formal manipulation of syntactically structured representations, or to put it slightly less polysyllabically, by moving symbols around according to rules. It's fairly plausible that this is true at least for some cognitive processes, but there is a wide scope for argument about whether this ability is the latest and most superficial abstract achievement of the brain, or something that plays an essential role in the engine room of thought.

Don't you think, to digress for a moment, that formal logic is consistently over-rated in these discussions? It enjoys tremendous intellectual prestige: associated for centuries with the near-holy name of Aristotle, its reputation as the ultimate crystallisation of rationality has been renewed in modern times by its close association with computers - yet its powers are actually feeble. Arithmetic is invoked regularly in everyday life, but no-one ever resorted to syllogisms or predicate calculus to help them make practical decisions. I think the truth is that logic is only one example of a much wider reasoning capacity which stems from our ability to recognise a variety of continuities and identities in the world, including causal ones.

Up to a point, Fodor might go along with this. The problem with formal logical operations, he says, is that they are concerned exclusively with local properties: if you've got the logical formula, you don't need to look elsewhere to determine its validity (in fact, you mustn't). But that's not the way much of cognition works: frequently the context is indispensable to judgements about beliefs. He quotes the example of judgements about simplicity: the same thought which complicates one theory simplifies another and you therefore can't decide whether hypothesis A is a complicating factor without considering facts external to the hypothesis: in fact, the wider global context. We need the faculty of global or abductive reasoning to get us out of the problem, but that's exactly what formal logic doesn't supply. We're back, in another form, with the problem of relevance, or in practical terms, the old demon of the

frame problem; how can a computer (or how do human beings) consider just the relevant facts without considering all the irrelevant ones first - if only to determine their relevance?

One strategy for dealing with this problem (other than ignoring it) is to hope that we can leave logic to do what logic does best, and supplement it with appropriate heuristic approaches: instead of deducing the answer we'll use efficient methods of searching around for it. The snag, says Fodor, is that you need to apply the appropriate heuristic approach, and deciding which it is requires the same grasp of relevance, the same abduction, which we were lacking in the first place.

Another promising-looking strategy would be a connectionist, neural network approach. After all, our problem comes from the need to reason globally, holistically if you like, and that is is often said to be a characteristic virtue of neural networks. But Fodor's contempt for connectionism knows few bounds; networks, he says, can't even deliver the classical logic that we had to begin with. In a network the properties of a node are determined entirely by its position within the network: it follows that nodes cannot retain symbolic identity and be recombined in different patterns, a basic requirement of the symbols in formal logic. Classical logic may not be able to answer the global question, but connectionism, in Fodor's eyes, doesn't get as far as being able to ask it.

It looks to me as if one avenue of escape is left open here: it seems to be Fodor's assumption that only single nodes of a network are available to do symbolic duty, but might it not be the case that particular patterns of connection and activation could play that role? You can't, by definition, have the same node in two different places: but you could have the same pattern realised in two different parts of a network. However, I think there might be other reasons to doubt whether connectionism is the answer. Perhaps, specifically, networks are just *too* holistic: we need to be able to bring in contextual factors to solve our problems, but only the right ones. Treating everything as relevant is just as bad as not recognising contextual factors at all.

Be that as it may, Fodor still has one further strategy to consider, of course - modularity. Instead of trying to develop an all-purpose cognitive machine which can deal with anything the world might throw at it, we might set up restricted modules which only deal with restricted domains. The module only gets fed certain kinds of thing to reason about: contextual issues become manageable because the context is restricted to the small domain, which can be exhaustively searched if necessary. Fodor, as he says, is an advocate of modules for certain cognitive purposes, but not 'massive modularity', the idea that all, or nearly all, mental functions can be constructed out of modules. For one thing, what mechanism can you use to decide what a given module should be 'fed' with? For some sensory functions, it may be relatively easy: you can just hard-wire various inputs from the eyes to your initial visual processing module; but for higher-level cognition something has to decide whether a given input representation is one of the right kind of inputs for module M1 or M2. Such a function cannot itself operate within a restricted domain (unless it too has an earlier function deciding what to feed to *it,* in which case an infinite regress looms); *it* has to deal with the global array of possible inputs: but in that case, as before, classical logic will not avail and once again we need the abductive reasoning which we haven't got.

In short, *'By all the signs, the cognitive mind is up to its ghostly ears in abduction. And we do not know how abduction works.'* I'm afraid that seems to be true.

Schmenomenal

12 February 2006

One of the key problems in the field of consciousness is the 'explanatory gap' between conscious experience (especially its subjective, phenomenal aspects) and the physical world. For some people this gap is so palpable and undeniable it constitutes in itself a strong Brentanoesque case for dualism: others, however, find it possible to deny, one way or another, that the gap exists at all.

One attractive strategy for dealing with the issue is to deny that there is any gap in reality - but affirm instead that the appearance of a gap arises because we think about the two 'realms' in different ways: specifically, that the concepts we apply to phenomenal experience are different, in some important and relevant way, from the concepts we use for the physical world. I say it's an attractive strategy, but of course, you never get something for nothing, and part of the price of the strategy is the slight weirdness of the idea of two radically different kinds of concept. If we normally used irreconcilably different kinds of concept for various different aspects of the world, this would come naturally - but we don't. We may well use different vocabularies for thinking about the same thing in different contexts; we certainly think of things on several different levels of explanation (my hand is part of the planetary ecosystem, part of my body, a collection of muscles, bones, skin, etc; and also a large collection of molecules, ions and what have you). But these different sets of concepts clearly have at least a limited translatability; it's unimaginable that one set or level could be disconnected and float off or disappear without affecting the others.

By contrast, if we want to adopt the conceptual strategy to get us over the explanatory gap, we have to embrace a strangely bifurcated conceptual world. To make that stick, we have to offer a convincing account of the differences and how they arise.

David Chalmers has produced a paper ("Phenomenal Concepts and the Explanatory Gap") which seeks to blow all such accounts out of the water. He mentions a number of variations on the conceptual strategy, and describes four particularly - namely the views that phenomenal concepts:

- are *recognitional*: they represent the way we think about things we've identified without having deployed our explicit knowledge about them;
- have a distinctive *role*, involving particular faculties or modes of reasoning;
- are indexical: words like "here" and "now" have a special relation to the speaker, and phenomenal concepts have a different but similar kind of special relation;
- phenomenal concepts are quotational: the concepts actually include the phenomenal state they have to do with.

These all have some intuitive appeal and face a variety of objections: but the details don't matter because Chalmers' argument is a general one designed to undercut any argument along these lines.

The argument goes like this. Concept theorists all say that there are certain special psychological features which human beings possess; that these features explain our epistemic

situation in respect of consciousness (in particular, why there seems to be an epistemic gap: a gap between the way we know some things and the way we know others); and that these special features are themselves capable of purely physical explanation. Chalmers argues that no set of psychological properties can comply with both requirements: either they fail to explain the epistemic gap, or they fail themselves to be capable of physical explanation.

Let C be the thesis that human beings actually have such properties. Chalmers asks: could there be zombies, creatures physically identical to us but without phenomenal experience, which *also* lacked the properties attributed to human beings by C? It looks as if we're heading back into the old familiar argument, but in fact this is where Chalmers has a neat move: it doesn't matter whether you answer yes or no.

If you think there *could* be physically identical zombie creatures which lacked C qualities, then C qualities clearly don't have a physical explanation - they can be there or not there and make no physical difference.

If, on the other hand, you think a zombie without C qualities is not conceivable, C qualities cannot explain our epistemic situation, and in particular, the gap. If any zombie has to have the same C qualities as its physical human analogue, C qualities can't be anything to do with the phenomenal experience that only real, non-zombie people have - nor anything to do with the perceived gap between that experience and ordinary objective knowledge.

If you accept the validity of this argument (and it looks pretty sound to me) it is pretty much a knock-down for any conceptual argument along these lines. Chalmers considers a number of possible responses, but the one I found most interesting was
the *schmenomenal* argument. This takes the line that, as a matter of fact, zombies do share our epistemic situation. It is argued above that zombies can't perceive the explanatory gap: but by specification their physical behaviour is the same as ours. It follows that they will at least talk about the explanatory gap: in fact they will write perceptive philosophical papers about it, in every way similar to those we might write ourselves. Surely being able to talk about something in a knowledgeable way is a very good test for one's epistemic situation? When the zombies talk about phenomenal experience, as they surely will, are they talking about nothing? They must, it seems, be talking about something else, which we might call schmenomenal experience.

Schmenomena seem to me potentially a whole new playground for philosophers. Speaking for myself, on the other hand, I don't see any room for them between phenomena and physical reality: either people ought to have schmenomenal experiences too, or more likely, zombies can't have them. But then I'm sceptical about 'ordinary' zombies in the first place. None of that matters so far as Chalmers' argument is concerned, of course: if schmenomenal experience is ruled out, that just bars the conceptualist's escape route even more securely.

A nice argument, then: but alas, it's also further confirmation for the view that all the best arguments in this field are refutations.

Memories

1 March 2006

Deborah Wearing's new book tells the story of her husband Clive's extreme memory loss and its consequences. Clive's amnesia, caused by a viral infection, is radical: with the exception of a few details, he remembers almost nothing about his life; moreover, he is incapable (again, with some minor qualifications) of generating new memories. The result is that he lives in a kind of perpetually disoriented present; always under the impression that he became conscious only about a minute ago; never knowing where he is or why, greeting his wife, about the only person he recognises, with effusive relief even if she only left the room a minute ago.

The question arises: how far is this man still the original Clive Wearing? More cruelly: how far is he still anyone at all? His wife says there is still "a Cliveness about Clive" , but is she right, and if so, is the "Cliveness" enough to constitute a person?

It has been argued, after all, that memory actually constitutes a person. John Locke, for one, seems to think so. Locke's view was that different criteria of identity applied to different kinds of thing. For rocks and other brute physical objects, simple physical identity was good enough. Living things were different: as the sapling grew into the mighty oak, or the colt into the full-grown horse, there were constant changes in the physical stuff involved; consequently the identity of plants and animals resided in the persistence of the same organised body, regardless of which molecules might be drawn into or expelled from that organisation. Presumably we have to allow a bit of latitude for the development of the organisation too: perhaps the sapling has the same organisation as the tree, but it's harder to spot the same structures in the acorn. The criteria of identity for a man, in any case, are the same as those for an animal: but criteria of identity for a *person* are another matter.

It might seem that Locke had a ready-made criterion of identity to hand, namely identity of the soul: but interestingly he rejects this on grounds which a modern materialist would find congenial. Since the soul has no physical aspect, a series of men from different ages might for all we know have the same soul, but it would be absurd to say they were the same man. If the soul of Heliogabalus entered a pig, it wouldn't make the pig a man – nor would it make it Heliogabalus.

So what is Locke's criterion of personal identity? Consciousness, in a word - more specifically, the kind of consciousness which binds together our different actions and experiences. Consciousness alone unites remote existences into one person. If we remember seeing some noteworthy event, we should never doubt that we are the same person as the person who had that experience. Memory thus becomes crucial to personal identity.

The distinction between the identity of human beings and the identity of people runs Locke into some difficulties. If a man genuinely has no memory of the crimes he committed while drunk, he is on Locke's account a different person and not, therefore, responsible. Locke bites this bullet quite firmly: according to him the law is concerned only with the identity of the man,

because for practical purposes that's all we can be sure of: but when God comes to judge us, perhaps we really will be forgiven for the things we did while drunk and don't remember.

If that's the case, then presumably our drunken selves, being separate persons, will have a separate existence (in Hell?) after death from our sober ones. What about the bad things we did immediately before getting drunk? I'm a different person from my drunken self, but we are both the same person as my pre-drinking self, so the two of us ought to be punished for what only one of us did. If there's compensation to be paid, the payee will get a windfall benefit, unless my drunk and sober versions are allowed to split the cost, in which case our liability for the earlier crime will have been halved by drinking too much.

What would Locke say about Clive Wearing? Clearly in Locke's terms the current Clive Wearing is the same *man* as the pre-virus version - but probably not the same person. However, it's a bit complicated. Clive still has a few memories: he remembers choosing which college to go to, for example. Does that mean he is the same person after all, or merely the same person as the Clive who existed at that particular moment, not the same as the one who went on to have further experiences, now forgotten, a month later (although that one was also the same person as the college-choosing Clive)? What about Clive now? To Locke, today's Clive is clearly a different person from yesterday's, about whom he remembers nothing. (In fact it's almost as if the Clive chatbot has been loaded up and run again: the same circumstances bring out the same jokes, questions, remarks.) But is the Clive of this present moment the same as the Clive of two minutes ago? Since he remembers nothing, he can't be: but there is an intervening Clive of one minute ago who he remembers, and who in turn remembers the two-minute-old Clive. If A is B and B is C, then… A is C? In fact, since there are no definite cut-off moments in Clive's recollection, his daily existence as a person has a puzzling kind of trailing edge about a minute or behind the present moment at any one time.

So would Locke be right about Clive Wearing? I don't actually think so. I have reservations about Locke's overall account of identity: I suspect it is over-sophisticated and does not draw a clear enough distinction between the qualities that distinguish a human being from a plant and those that distinguish one human being from another.

After all, is simple physical identity good enough even for rocks? If a piece, or two pieces, are chipped off a rock, does it really take on a new identity? How are we to deal with the identity of rivers, since as Heraclitus famously pointed out, you never step into the same, physically identical, river twice? We need to allow the identity even of simple things to depend on a looser concept of general physical continuity, but once we do we find that physical continuity works pretty well for plants, animals, and dare I say it, human beings too. It may well be that memory is a key human capacity – it's certainly hard to imagine that someone deprived of even the limited faculty that Clive possesses could function mentally at all. Nevertheless, it seem quite possible to me in theory (though unlikely in practice) that two people could have identical memories, and I don't at all see that that would lead to their losing their separate identities.

In short, I think Clive's simple physical continuity guarantees his continued identity as both a man and a person: he just has a memory problem. A really bad memory problem.

Really Selfish

10 March 2006

Maybe genes really are selfish, says Stephen Pinker, in a short essay published recently in the Times. The piece is one of a collection which OUP will be publishing as a tribute to Richard Dawkins, though the idea that the selfishness of genes might be more than just a metaphor seems a slightly ambivalent choice of theme in this context. As Pinker points out, since coining the phrase 30 years ago (yes, it really is that long) Dawkins has spent a fair amount of time carefully and repeatedly explaining that no, he doesn't suppose genes have tiny brains and personalities, it's simply a metaphor, albeit a vivid and useful one. Dawkins has reacted good-humouredly in the past when some of his ideas were taken rather further than he might have wanted to go with them himself – notably in the case of Susan Blackmore's bold extension of the idea of memes into a full-blown theory of consciousness. But Pinker's suggestion surely puts him in a difficult position: he either has to disown the friendly tribute and seem ungrateful or risk looking inconsistent.

What is Pinker on about anyway? He praises Dawkins for seeing life and evolution as matters best understood though ideas about information and computation and suggests a similar perspective has shaped modern views about cognition. Another common feature between Dawkinsian biology and recent cognitive science, he suggests, is the use of a 'mentalistic' vocabulary. That rather glosses over the point that cognitive scientists try to keep mentalistic terms out of the business end of their theories – if an explanation of mentalistic concepts uses mentalistic concepts itself, we obviously still have work to do.

It also picks on the very feature of Dawkins that I personally like least. OK, so the selfishness of genes is a metaphor. But a good one? If we are to think of ourselves as lumbering gene carriers, we have to assume that the passengers designed a craft which randomly minces some of them every time they change ship and seeks a better class of passenger for the next generation rather than offering its own original genes continued safety. I wouldn't build a vessel like that for myself. But worse, the metaphor hides from view the cardinal virtue of Darwin's theory, namely that it does away with the need to invoke intentions and conscious design: that elegant economy of means is the prime reason why Darwin's version of evolution is better than all the others. It's as though Dawkins were to explain Newton to us in terms of angels wanting to push masses together: altogether better, if you can, to make the small effort required to describe things the way the theory says they actually are.

Pinker, however, argues that we can safely - in fact correctly - use mentalistic terms about genes and the camouflage patterns of animals (which embody 'knowledge' about the environments of the animals' ancestors). After all, when we use such terms about human beings, they aren't about phenomenal conscious experience, because nobody understands that anyway: they're about the other, computational stuff, and all that is an 'entirely tractable' scientific topic.

That seems rather breezy. Of course there are those who think intentionality, meaningfulness, can – ultimately - be reduced to computation or information processing theory, but we haven't yet reached the stage where that view is uncontroversial: and indeed, it isn't clear to me in any

case that all such theories abolish the distinction between the ‘design’ of a leopard and the design of a battleship. In my own view, the core mystery of intentionality is just about as intractable as the mystery of qualia.

In fairness, it is true that it is sometimes hard to talk about complex organisms without relapsing into ‘mentalistic’ or teleological terms: it’s hard to say that when a baby bird hatches, its wings don’t embody some kind of expectation of flying. But I would argue we need a new vocabulary for this kind of thing, along the lines of the ‘natural meaning’ of H.P.Grice (‘those spots mean measles’). Indeed, on another occasion I’ve suggested we could co-opt the word ‘point’, conflating two of its senses: the point of wings is flying, and wings point to flight. But while I sympathise with Pinker’s thesis to that extent, I think it would be confusing and unhelpful to apply ordinary ‘mentalistic’ terms where they don’t strictly belong in anything other than a metaphorical sense.

Of course, it is true that in ordinary usage we do attribute intentions and design to natural objects – but not necessarily in the way Pinker would want. He, I think, would expect us to talk about how the leopard’s spots derive from the knowledge it, or its genes, have: but it would be more normal to talk about how perfectly designed for its lifestyle the creature is, or how well Providence has equipped it for its role. I can see the proponents of Intelligent Design patting Pinker on the back: “That’s right Steve: of course we should talk about the aims of organisms and the thinking behind their design: we won’t say it’s G__, but we agree with you it’s more than a metaphor…”

The Mereological Fallacy

21 March 2006

It isn't actually your brain that does the thinking at all. In fact, the very idea that it does is virtually incoherent: not just wrong, but meaningless: the *mereological fallacy*. Only you as a whole entity can do anything like thinking or believing.

That's pretty much the bombshell dropped by Max Bennett and Peter Hacker in their 2003 blockbuster *Philosophical Foundations of Neuroscience*. Mereology is a branch of logic dealing with the relations of parts and wholes; it represents one of the attempts which were made to sort out the chaos left behind in the wake of the catastrophic collapse of Frege's theories. Bennett and Hacker's book has given the term a new lease of life with a somewhat looser sense. Actually, the book is large and complex, and I can't do anything like justice to it here, but the idea of the mereological fallacy, a kind of leitmotiv running through the book, certainly deserves some attention.

What do they mean? If I don't think with my brain, do I think with my foot? There is, of course, a sense in which my foot seems to know things. It 'knows' roughly what forces to apply where in order to keep me walking straight. My arm knows how to move in order to be in the right place to catch a ball heading in my direction, not a process I myself could describe in any great detail. But although athletes talk about muscle memory, this kind of knowledge is really little more than a metaphor, or at best an example of knowing how to x, not of knowing *that* x. It certainly isn't what Bennett and Hacker have in mind, anyway: the last thing they want to do is transfer cognitive functions to another mere part of the body: their point is that only whole people have these capacities.

But what's a whole person? If my foot, or even my leg, is cut off, it doesn't seem to have any relevance to most of my thought processes. Indeed, if the stock science fiction/philosophy example is to be believed, a brain removed from the body altogether and sustained in some kind of life-support tank could perfectly well go on thinking for itself. That may be no more than a thought-experiment with a dubious basis in medical reality, but it does suggest that the brain is the thinking organ, just as the heart is the blood-pumping organ.

I think that to appreciate Bennett and Hacker's point properly, you have to consider it in historical context (they provide plenty of this: I'm afraid poor Descartes gets yet another drubbing for inflicting his dualism on us all, though for once his positive contribution is also acknowledged). Philosophically, I think we can see the doctrine of the mereological fallacy as being in opposition to the problematic old theory of sense-data. For a very long time, it was accepted that we didn't see the world, we saw sense-data (or images, or some similar intermediary). In philosophy, this view was popular because it seemed to make it easier to explain cases of error or illusion; it was just defective sense-data. Scientifically the fact that a nice image appeared on the retina, *within* the eye, must also have predisposed people in its favour. If discussions of the senses had been mainly about touch, where there is no apparent intermediary, rather than sight, things might have been different.

On the whole, I think it has been accepted within philosophy that the sense-data point of view was a confusing mistake: when we see a table, we see a table, not some patches of brown

within the visual field: the brown patches, the images and the sense-data, if there are such things, are just part of the means by which we see. Bennett and Hacker, if I understand them correctly, want a root and branch application of this kind of revised thinking to all aspects of cognition. Brains sustain patterns of neuron firing, and patterns of neuron firing correlate with cognitive activity: but to say that the firing neurons *are* the thoughts is just a mistake. As a bonus, this (apparently) eliminates the binding problem: if your experiences aren't in your brain at all, the question of where in your brain the different elements are put together no longer arises. (But perhaps some related questions still do!)

Up to a point, this is quite right, and a necessary corrective. I think Bennett and Hacker are probably right to think that Crick, for example, has fallen into the trap they are describing: but it's not so clear to me that everyone has. Some people, and Crick was one of them, do assert that neurons firing just *are* mental experiences, full stop: but most people, even those who assert the identity of mental and neural events, only mean that the two things are different aspects of the same phenomenon; the same, but on different levels of interpretation.

Perhaps an analogy will help to make this clearer. Hennett and Backer, let's say, have a controversial theory about phone conversations: when you call your mother, they say, the conversation does not go through the local telephone exchange: it can't be found in some cable somewhere: that's just nonsensical - it takes place *between you and your mother*. Now the theory may seem a little unexpected, but the thing is, the opposite view has held sway for a long time. People are always saying things like "You imagine you're talking to your mother, but actually science has conclusively established that you're actually talking to the phone. OK, the phone may tell you something about what your mother is saying – maybe quite a lot – but it's only the phone you're talking to. You actually have no contact with your mother, and no certainty about what she's telling you."

In this case it's obvious that there are two equally valid levels of interpretation: it's just that the two sides are being pig-headed about recognising it. Of course you're talking to your mother, and having a conversation which we wouldn't normally describe as happening in a cable: but in another sense you are addressing the phone (your mother's miles away!) and there is a sense in which the conversation passes though the local exchange. If we imagine the telephone exchange as very old-fashioned (and I do) it's actually crucial that the operator knows which plug and socket your words need to travel down.

So yes, it is people who see things and think thoughts: but so long as we keep hold of that fact, and don't allow ourselves to be seduced by the sterile charms of eliminative reductionism, surely we can safely talk in another valid sense, about the brain doing these things?

Onphenes

1 April 2006

Ted Honderich's new theory of consciousness – that me being conscious of a room is in some sense just there being a room – was mentioned in these pages a while ago. On that occasion it was noted that, in spite of his philosophical eminence as a former Grote professor, his paper had been rejected by the Journal of Consciousness Studies (not normally a narrow-minded periodical - happy to publish pieces on parapsychology, say or Rupert Sheldrake), and it was suggested that the reason might be that no-one could work out what he was on about. But we liked the rallying-cry with which he ended, calling for further progress.

That call has not gone unanswered. Last August in Copenhagen, at Towards a Science of Consciousness, Honderich, Riccardo Manzotti and Francois Tonneau presented papers in support of their own versions of the theory of Radical Externalism. We could be witnessing the beginning of a movement. Suitably abashed, the JCS has apparently agreed that an entire issue will be devoted to responses to Honderich, some time in the summer.

The least we can do, then, is to have a further look at what these Radical Externalists are saying, and Manzotti's paper seems a good place to start. Manzotti very kindly responded to the initial appearance of this piece with his own comments – I have included these where relevant labelled 'RM'.

The distinctive feature of Manzotti's theory is that he takes a process-based view. Things don't exist, in his eyes, they take place. Like many others, he thinks we are to some extent captives of the philosophical outlook adopted four or five hundred years ago, but instead of Descartes getting all the blame, as is usual in these cases, he attributes the main guilt to Galileo (though of course Descartes does not escape unreproved altogether). Galileo, he says, adopted a methodology which separated the observer from an observed world of autonomous objects susceptible to measurement and mathematics. Some features of these objects, such as mass, were really out there in the world; others, notably colour, were really only in the observer's head. This, clearly, was a productive approach: but what Galileo had no business to do was to add an ontological commitment. The fact that considering the observer and the observed separately allowed him to do some interesting sums did not mean that the observer and the observed were really separate.

This mistake, according to Manzotti, led to our disastrously dualistic outlook, the view that perception happens in the head, and all our difficulties with connecting the inner subjective world with the physical reality outside. Manzotti is surely right to want to remove unnecessary intermediaries from our account of perception – what he refers to as the 'television' view.

In fact, he says, the mind is identical with everything the subject is conscious of. Instead of talking about the perceived and the perceiver, we should talk about the unifying process of perception, the 'onphene'

RM I do agree. However I would like to stress the fact that when I say that the mind is identical with everything the subject is conscious of, I refer to a world made of processes and not objects. I would not say that when I am aware of a bottle (as an object), I am identical with the bottle (as an object). I mean that the bottle is a process and in that sense I am identical with it.

. Our mental life is composed of these onphenes, reaching out far beyond our skulls. Once we adopt this conception of 'the enlarged mind', all the difficult problems of consciousness fall away. We don't have to reconcile qualia with objective reality because they are objective reality; we need no longer puzzle over intentionality, the link between things which mean and things which are meant: the onphene is intentionality.

According to Manzotti, perception in some sense constitutes objects.

RM Again. I do agree. Of course the processes, which constitute my perception, are also constitutive of the objects I perceive - the two being two ways to look at the same process. But let me stress the fact that I do not assign any particular power to the perceiver's side.

A favourite example is the rainbow. There is, in fact, no vast coloured arch in the sky: it's only in the act of observation that the rainbow is brought about. As with Honderich, it would be possible to misconstrue this as a radical kind of relativism – your own perceptions are the only reality – but Manzotti's outlook is much more commonsensical than that. Rainbows, after all, still have some underlying physical support in the form of raindrops and sunlight. I fear this is more commonsensical than it really has any right to be. If there's still an objective physical world underlying and giving rise to the onphenes, I think some of the old problems of the relationship between mental and physical are merely going to be relocated. If Manzotti wants to embrace his radicalism fully, I think he is bound to adopt an ontology in which there really is nothing but collections of onphenes.

RM As I mentioned at the beginning I would really like to opt for an onphene-only ontology. However we do not need to get rid of the familiar world of molecules and such. There can be a dual perspective on things. More or less like referring to the south and north poles of magnetic fields. However, you're right. This is a point that needs further clarification.

He would then be faced with difficult problems of accounting for why the onphenes exist, and how they come to have certain consistencies and commonalities – the kind of regularities which real objects at one end, and observing people at the other, are generally thought to explain.

RM I do not believe that those problems will be so difficult. However, I would like to be challenged by precise reference to cases that do not seem to fit with the process view. I have always had problems in dealing explicitly with them because usually philosophers refer in a vague way to them. I would like very much to do it.

One of the challenges he faces, of course is how to account for erroneous perceptions. If all my perceptions just are the things perceived, how could I ever suffer from illusions or mistakes? Manzotti offers an account of dreams and memories in which he seeks to preserve the idea that they are, somehow, realities. Memories are just delayed effects of the onphene still doing its stuff – but hang on there: if my mind is identical with the things it is conscious of, how come my mind is here now and the thing remembered is back then? Dreams, on the other hand, recombine elements drawn ultimately from reality. It is a key point for Manzotti that the contents of the mind can only come from the real world – you cannot imagine anything entirely new, such as a new colour. If I dream of my mother wearing Napoleon's hat, it is my real mother and (by a

more obscure process, since I have never seen it) Napoleon's real hat that are involved. But surely my-mother-in-the-hat is a legitimate single object of perception, not irreducibly dual. What if I dream of a purple cow? It may be, at some remove, a real cow: but where does the purple come from? A particular purple object? Purpleness?

RM This is a crucial issue for me. All experience must be rooted in reality. I think there is plenty of evidence for this. People who are born blind do not dream of colours. We do not dream ultrawaves of infrared, and so on. As far as we know all conscious experience is rooted in reality. During perception processes have a short time span. In memory the time span is longer. Where is the difference? Could I dream of my grandmother unless my grandmother had been there many years ago? Your example does not quite reflect my view correctly. According to my position you cannot dream of Napoleon's hat, since that hat has never been part of your experience. There is no causal connection between that hat and you. But of course you had plenty of direct contact with paintings or other kind of reproduction of Napoleon's hat. Thus there are two events: one is your grandmother at a certain point of your life and another is a reproduction of Napoleon's hat at some other point in your life. These events(where "event" is a way to refer to the beginning of a process) will eventually connect (probably due to some random links between your neurons). Because of this kind of joint effect the two processes intermingle in one. Still, both the grandmother and the reproduction of Napoleon's hat, by means of what takes place in your skull, are something different. In some respect, they do exist thanks to that process.

Let's get to the purple cow. Where does the purple come from? Every time we perceive a purple object in our V4 there is the perception of purple. What is purple? It is a certain cause (a certain ratio among the frequency of light waves) singled out by certain process (those ending in V4).

Manzotti goes so far as to claim that the coloured spots created by pressing your eyes (or banging your head – don't try this at home) derive from real colours we have seen. A person who was blind from birth whose eyes were pressed in this way, he thinks, could not experience colour, but would perceive the spots through the sensation of touch (why not smell?).

RM Because there is a causal connection between tactile events and the brain by means of his/her blind eyes. The eyes, presumably, are subject to tactile pressure and thus there is a causal connection between pressures and the further neural activity. This can be tested by pressing the eyeballs. However, it could be possible to envisage a more complex experimental setup in which there is a transducer of the density of certain molecules to the neural activity on the optic nerve. In this case, I suggest, there will be an olfactory sensation by means of the optical nerve. What counts, in my view, is what is brought together by the causal process, not the machinery in between.

But the weakest of his defences, I think, is the one presented for the case of optical illusions such as the Kanisza triangle (formed by three 'pac-men' providing the apparent corners) or illusions of movement. Manzotti introduces the idea of a 'perceived triangle' and 'perceived motion' as the objects of perception in these cases. Wasn't this just the sort of thing he was against? And where is the perceived triangle? Not, surely… in the head of the perceiver?

RM The weakest of my defences! I need to work harder on it then! Again I did not express myself in the best way with the use of terms like "perceived X". Well, I won't be very precise on this issue. The general idea is the following: there are processes that single out certain combinations

of events. Usually these processes take place together with other processes. Sometimes these associations are disrupted. And yet there is a real continuity with the first kind of process.

In short, I think there are unresolved problems in Manzotti's theory. He might well seek to resolve them by becoming less radical about his externalism – but perhaps it would be more interesting to go the other way: cast off from the shores of mere common sense and set up a truly radical onphene metaphysics.

RM I think there are yet many unresolved problems and I like very much what you suggest to me at the end of your page! I hope so.

I seldom find such a clear and to-the-point understanding of my work. For this reason I would like to try to dispel a few minor misunderstandings that might be attributed to my paper.

In my paper I was, due to many suggestions on the part of the editors, rather conservative and, as you put it, maybe a little more commonsensical than I should have been. The reason is that, when you present a really novel viewpoint, readers seem to have problems in following you unless you keep a low profile.

Let me give you a very sketchy but extremely honest summary of my view.

According to the traditional view, reality is made of entities that are under many respects autonomous and separate. Among the proposed entities I could quote atoms, objects, events, individuals and many others. The problem is that, once you accept such a view, reality is made of separate entities that cannot be put together again. The problem becomes severe in the case of conscious experience whereas a body (or a brain) has an experience of something else (the external world). How can the two be one if they are made of separate entities?

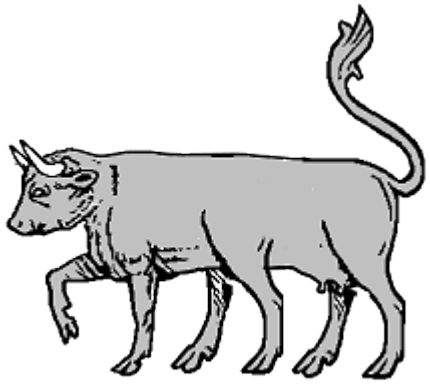
I suggest accepting an alternative view: reality is made of processes. Each process is like a dipole - it can be seen from two perspectives (cause and effect) but it is really one process. Whenever we refer to an object or a physical state of affairs we refer to a process taking place.

Unlike objects, processes are compositional. I mean they merge together forming other processes.

Thus, consciousness becomes an opportunity to understand the nature of reality. It is not another problem. It is a chance to grasp the overall nature of things.

Paths of Consciousness

22 April 2006



What about Francois Tonneau, then, the other radical externalist who has rallied to Honderich's side? In his paper he actually positions himself as reviving and improving the stance which was apparently taken by psychological neorealists ninety years ago.

The bedrock of Tonneau's externalism seems to me to be the same intuition which inspires Honderich and Manzotti: as he says near the beginning of the paper¶" if consciousness is a brain process, then conscious experience cannot have the features that in fact it has, for these features are features of the environment (such as the colours and shapes of surrounding objects)."

That seems to me to confuse features of experience with properties of experience: my experiences don't have to be red in order to feature red objects. Tonneau, approvingly, quotes Holt decrying the idea of shapeless representations of shape and colourless representations of colour –

"my knowledge is neither shapeless, motionless, colourless, or odourless".

Well, I rather think it is, actually. But the force of the argument lies mainly in its appeal to the evident difference between riotous phenomenal experience and greyly firing neurons.

Tonneau offers two distinctive ideas. The first, picked up from the earlier theorists he mentions, is that of a cross-section. A cross section is (deep breath) a function of the state of a reference system that takes its values in the environment, the value at any moment being the content. Tonneau very helpfully likens all this to a torch whose beam can pick out various objects depending on which way it is turned: the beam is like the cross-section with the content being whatever object in the environment happens to be illuminated at any one time.

This idea allows Tonneau to deal with a number of objections to externalism. If our consciousness is external, why do changes in our brain affect it, as they clearly do? No problem – for Tonneau, it's simply as though the torch were being moved around: the change is internal, but it changes the content of something going on out there in the environment, as the beam highlights a different object. In much the same way, it's easy to explain how people can go on having conscious thoughts while the external world, where their consciousness resides, is static.

You might feel that this line of thinking dilutes Tonneau's externalism, because it implies that something very important, something which helps determine the contents of consciousness, is actually going on in the reference system – which must surely be on the inside? Personally, I think the proposed set-up works quite nicely except that it incurs the debt of having to assume that all the potential objects of consciousness are actually already out there in the environment.

Tonneau's second idea helps deal with this commitment, and also provides a way of addressing some even more serious objections: if conscious experience is external, how can we ever dream, or imagine things, or suffer hallucinations? The idea here is that phenomenal properties are higher order features of a person's path. A path in this sense is the whole series of

experiences gone through in the course of the person's life. So when we imagine something, we are drawing on elements which were in our environment at some point; they may be recombined in a way which gives the appearance of originality and novelty, but they are all drawn from reality. This explains how we manage to think about things we are not experiencing, and it goes some way to explaining how there comes to be such a huge range of potential objects of consciousness apparently just available in the environment.

Not so fast, you may well think: if my conscious experience is external to me, its objects must be around now, mustn't they? But these objects from my path are all back in the past. You can't experience things which are no longer present, and indeed may no longer exist. Tonneau's tactic here (if I've understood it right) is connected with the claim that phenomenal qualities are higher order properties. Suppose I saw a blue object last week: the object had the property of being blue, and so my path has the property of containing an object with blue properties. According to Tonneau, that property of containing-an-object-with-blue-properties itself has a property which is the thing I experience now. In fact, that property is the colour we actually experience. Basically, Tonneau uses the temporally-extended nature of the path as a kind of bridge to bring things forward out of the past for us to experience.

This all looks a bit fishy. It seems an odd idea that the colour I see is actually something as abstract as a higher-order property of the totality of my life experiences. I'm not sure the bridge effect really works, either. Actually only part of my path has the property of containing a blue object, and it isn't the part that's here now. England contains many churches, but that doesn't mean that if I can see any part of England I can see churches.

I'm also unconvinced by the idea that the imagination works merely by re-shuffling components derived from our experience. If I think of a purple, six-legged cow, Tonneau would say I've put it together from a cow I once saw and past experiences of purpleness, simply adding an extra set of legs. But if I have a purple bovine hexapod in my conscious mind, I don't just have a cow, a colour, and some legs: I also have the purple, six-legged cow: the theory that consciousness is external surely requires that this legitimate object of thought must itself be out in the environment, not just its separate components? Tonneau believes that his system makes it possible for us to suffer illusions and other non-veridical perceptions, but I don't think he has succeeded – without a gap between the world and our consciousness, I don't really see how our perceptions can ever be mistaken.

At times when reading the radical externalists' arguments, I wonder who they are arguing against. They believe that the majority view is internalist, believing that consciousness resides in the brain, but surely very few people believe that in the straightforward sense. Where is the story of Macbeth? In one sense, in Scotland: in another, in Shakespeare's manuscript; on stage, in our minds, even in our brains. But in some final, metaphysical sense, stories don't have a physical location at all, and neither does consciousness. The chief problem with externalism is not so much that it puts consciousness in the wrong place, as that it insists on a location at all.

This point of view is explicitly tackled by Tonneau, along with a number of other objections, but he merely offers reasons why we might be inclined to think our experiences have no location, a response which falls well short of a refutation.

But there's more to be said yet

Anaesthetic Entanglement

5 May 2006

Huping Hu and Maoxin Wu's new paper "Photon Induced Non-local Effects of General Anaesthetics on the Brain" moves things on rather dramatically. You may recall that earlier papers invoked quantum entanglement as an essential mechanism in the operation of the mind. The new paper says that, rather than waiting for things to happen, they decided to go ahead and conduct some experiments. A range of slightly different tests was carried out: it appears that applying magnetic pulses through a sample of anaesthetic into the brain causes distinct effects as though the anaesthetic were actually in the brain: moreover, water can be made to have anaesthetic effects through a similar procedure: the effects of drugs, in fact, can be transmitted into the brains of subjects through quantum entanglement.

These experiments rather recall the homeopathic doctrine of water memory (hopelessly at odds with elementary chemistry, in my personal view), but the authors distance themselves from that theory. More frivolously, they reminded me of Bunuel's remark – "Connoisseurs who like their martinis very dry suggest simply allowing a ray of sunlight to shine through a bottle of Noilly Prat before it hits the bottle of gin."

In Hu's own words the significance of these results is as follows.

i) ***Significances in Physical Sciences***

1. Our findings enable communications of both classical and quantum information between locations of arbitrary distances using quantum entanglement alone.
2. Our findings show that instantaneous signalling is physically real which implies that Einstein's theory of relativity is in real (not just superficial) conflict with quantum theory.
3. Our findings provide important new insights into the essence and implications of the mysterious quantum entanglement.
4. Our findings further provide clues for solving the long-standing measurement problem in quantum theory including the roles of the observer and/or consciousness.

ii) ***Significances in Biological Sciences***

5. Our findings show that biologically and chemically meaningful information can be transmitted from one place to another by photons and possibly other quantum objects such as electrons, atoms and even molecules through quantum entanglement.
6. Our findings enable various quantum entanglement technologies be developed some of which can be used to deliver the therapeutic, nutritional and/or recreational effects of many drugs to various biological systems such as human bodies either on site or from remote locations of arbitrary distances without physically administrating the same to the said systems.
7. Our findings suggest that brain processes such as perception and other biological processes likely involve quantum information and nuclear and/or electronic spins may play important roles in these processes.
8. Very importantly, our findings provide a unified scientific framework for explaining many paranormal and/or anomalous effects such as telepathy, telekinesis and homeopathy, if they do indeed exist, thus transforming these

paranormal and/or anomalous effects into the domains of conventional sciences.

The results are certainly surprising. I don't think I envisaged the effects of entanglement being quite along these lines, as though a distant molecule could have its normal chemical effect in another location. But the main barrier to acceptance of the results, I think, will be the nature of the experiments themselves.

Rather than being conducted in a well-equipped lab, they seem to have involved a degree of improvisation, using an audio system, a microwave oven, and a flashlight, among other things. Some of the drugs were "leftover items originally prescribed to Subject C's late mother". None of that invalidates the results, but the mention of Subject C indicates another issue: I think readers will feel that there are methodological problems in the fact that the four subjects of the experiments were the two authors and Hu's parents. I'm afraid this is likely to deter others from attempting to reproduce the results, and the experiments may not enhance the credibility of the earlier paper in the way the authors presumably hope.

[Note: Hu and Wu had faith in their own results and apparently applied for a patent on this method of administering drugs, which was rejected on the grounds that it did not work.]

The Deepest Woo-woo

7 May 2006

Galen Strawson's paper "Realistic Monism: why Physicalism entails Panpsychism" (to be the keynote of "Consciousness and its Place in Nature", due out in September) has already attracted favourable attention. It's certainly an amusing read – though the entertaining tone perhaps includes a hint of bluster here and there. In any case my own view is that it contains some dreadful pieces of argument, and in no way delivers what the title promises. It promises much, of course: that physicalism entails panpsychism? That merely by sitting and thinking quietly about the nature of physicalism we shall see that everything is or has a soul? Establishing that would indeed be a magnificent achievement. There are three small snags: Strawson doesn't mean "physicalism", he doesn't mean "panpsychism" – and he doesn't mean "entails".

Strawson makes it clear that when he speaks of physicalism, he doesn't mean what most people would mean: no, he means real physicalism, which acknowledges that experiences are physical, and that they are all we really know of in the world, the necessary starting point for any enquiry. Of course the objects and indeed the subjects of experience are often undoubtedly physical, but it's an odd idea that experiences themselves are concrete physical objects with a definite spatial location. I'd like to see Strawson point to some of his sometime, or catch some in a jar to use as evidence.

He distinguishes true physicalism from physicSalism – the belief, utterly false, he says, that all of concrete reality can be captured in terms of physics. But every real, concrete phenomenon is physical according to Strawson: he must therefore believe that there are physical phenomena which lie outside the scope of physics. Physical phenomena which lie outside the scope of physics? What could those be? Never mind whether this is good philosophy - it isn't even good lexicography. Perhaps it's just that Strawson's exposition has gotinto a bit of a muddle. Rather than giving a radically different sense to the word "physicalism", it would be better to choose another term – as he partially acknowledges. He suggests we could call his theory "experiential and non-experiential monism", but I think "experientialism" would be a fair description.

What about Strawson's alleged panpsychism? He doesn't seem, in fact, to believe that souls are everywhere: only that experience is the ultimate building block of the world, which we could better call panexperientialism. So the proud boast of the title is reduced to the claim that "experientialism entails panexperientialism". How does the entailment work? Stripping away many twists and turns, we come down to the observation that unless everything is constituted by experience, there must be some other stuff around, and "I would bet a lot against there being such radical heterogeneity at the very bottom of things". Well, here's a fiver says you're wrong, matey.

You're entirely missing the whole point. First of all, Strawson is absolutely right to insist on the primacy of experience. People often suppose that science is about our real experience of the world, the things we know about for sure, while phiosophy is all about complex abstractions

which we know of only through complicated pieces of reasoning. In fact, of course, it's the other way round: phenomenology deals with the undeniable realities of what we experience, while the entities described by physics have to be inferred from those experiences - what else have we got to go on?

But what you're wilfully ignoring is Strawson's main argument. His point is that if experience is not fundamental, it must have arisen out of non-experiential stuff. Now he accepts the phenomenon of emergence, where in particular circumstances a particular arrangement or organisation of a substance can have properties which the substance in itself does not have. But, he points out, emergence cannot be brute: you can't get new properties emerging at random. There has to be an intelligible account of how the emergent properties are constituted by the properties of the substrate. In the case of experience, it is clearly impossible to give any account of how experience could be constituted out of a non-experiential substrate: so experience must be fundamental. It might be the case that there is also a non-experiential substrate, but it's more economical to assume that at a micro level the world is constituted entirely from experience. Put these micro-experiences together in one way, and you get a non-experiential rock, or table: put them together another way and you get a higher-level experiential mind.

I'm not sure that's actually right, but it's an appealing scheme.


I'm not ignoring anything: there is no argument to ignore. Strawson wants to resurrect the medieval principle that ex nihilo nihil fit - nothing will come of nothing. Five hundred years ago, he would have been saying that that proved our minds could only have come from the greater mind of God: now he wants to say they could only have come from a world full of micro-minds. But the principle is just wrong, or at least subject to the very large exception of emergence, which does allow something to come out of nothing after all.

Of course there should be an intelligible account of how the emergent phenomenon emerges: but the fact that we haven't got that account for consciousness yet doesn't prove it's impossible. Virtually the whole subject of consciousness studies is an attempt to give an account of how experience emerges from non-experience. To assume there cannot be such an account just begs the question. What Strawson really wants to do is deny the possibility of emergence; but he knows that is totally implausible, so he behaves like an amateur conjuror - he talks around the subject, moves it around a bit, shifts it back, talks about something else, and then suddenly emergence disappears up his sleeve. I think I can point to the moment when it happens - he's just been drawing our attention away with some talk about "Z properties", and then all at once:

"For what we do, when we give a satisfactory account of how liquidity emerges from non-liquidity, is show that there aren't really any new properties involved at all."

There is no emergence, all explanations of emergent phenomena are eliminative, and ex nihilo nihil fit! Neither true nor supported by the preceding discussion.


Yes, all very rhetorical - but don't you have to concede the primacy of experience, as I said before? And if experience is the only direct reality we know about, isn't it a strong candidate for the basic building block of reality?

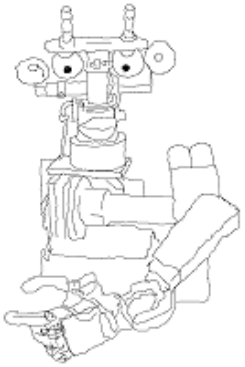
 No: for two reasons. Strawson spends a lot of time waving his hands over how his opponents - he names Dennett, for example - deny the reality of experience. This “deepest woo-woo of the human mind” is to the lasting shame of philosophy as a whole, it seems. But I ask - isn’t Strawson weirdly denying experience himself? He talks about experience as though it were an isolated phenomenon, but in fact the strongest feature of experience is that it is about something, and in fact, about physical reality. We don’t just have experiences and then in our quieter moments think that, hey, these experiences could be clues to an external reality: we actually experience the world and then, in our quiet times, deduce that we know about the world through something we could label experience. Strawson talks as though experience came first, and then the tenuous hypothesis of a physical world came along later: but that actually denies the most salient features of experience.

In the second place, he chooses not to draw an important distinction between our experiences and those of other people. It’s only our own experiences we have first-hand, directly: from them we infer the details of the world around us, and then another step is required to believe that other entities out there are having experiences like our own. If Strawson were rigorous, he would conclude, not that the world is probably constructed out of experience in general, but that it is probably constructed out of his own experience: he would be a solipsist. Building our own experience out of the experiences of other entities is no simpler than building it out of non-experiential entities - perhaps a good deal less simple.

But what a feeble attitude anyway! Reality must be constructed out of the things I know best, as though the basic components of the world were human beings! The truth is that physics is indeed all about phenomenal experience: it simply provides a sophisticated and well-tested set of theories about the underlying, independent realities. To ignore all this and base your conception of the world on primitive experience is perverse.

Where do I begin?

22 May 2006



l was interested recently to read about Babybot, a research robot intended to model some of the characteristics of a two-year old child. Babybot reminded me slightly of Steve Grand's Lucy without her mask (there seems to be a consensus in engineering circles that for consciousness you only need one arm and no legs). A bad omen, I'm afraid: poor Lucy has apparently been gathering cobwebs for a while now.

The thinking behind Babybot is based on a process model of consciousness, which sounds interesting, but my impression is that the researchers have spent more time on the technological challenges of the sensorimotor apparatus than on the philosophical issues (quite reasonably, no doubt).

It wasn't so much that that interested me, though (and provoked a largely unrelated chain of thought), as the idea that you needed to produce a baby's consciousness before moving on to the adult version. As a practical research strategy, this has some obvious appeal - infant movements and senses provide a slightly easier challenge and may yield insights into the developmental process. But could it be that there is actually a stronger constraint here - that consciousness cannot be generated full-blown, but has to go through embryonic and infantile forms? Alan Turing certainly implied that this was a possibility in his famous paper of 1950, albeit in a tone which characteristically mingled the frivolous with the profound ("It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle."), and even said that he had conducted some experiments on a child machine.

The emergence of consciousness in human beings is itself an unclear and controversial matter, of course. We can feel pretty sure that a newly formed zygote lacks consciousness (perhaps I ought to specify human-style consciousness, for those of a panpsychist leaning); we can feel reasonably sure that a two-year-old has consciousness, though perhaps without the refined self-awareness of an adult. But we don't know exactly when consciousness dawns, and we don't know whether it is like switching on the light or something much more gradual, passing through a series of partly conscious states (whatever those might be). I think we tend to assume that the arrival of consciousness could be sudden, even if it isn't: that in principle we could construct an artificial consciousness in any arbitrary state x, where x might correspond with say, thinking about the cup of tea you're going to have when you get home, or trying to remember what your brother gave you for your twelfth birthday.

But not all states of affairs are programmable, and it might be that conscious mental states are not. Even machines sometimes need to be constructed in a particular sequence, so that state z, the finished product, can only be reached through a suitable series of earlier states. When in operation, some machines also have states which are constrained by sequences of previous states. Analogue clocks provide a simple example: you can't get the hands to a reading of tea-time without passing through readings which correspond to adjacent times. I once saw an orrery, which showed not only the date and year, but the position of the planets at any given time. Most clocks allow the hands to be disengaged from the mechanism and turned quickly to any time - a good enough practical approximation to being able to set times arbitrarily: but in this one the mechanism did not allow the planetary 'hands' to be wound forward quickly. If the

clock ever stopped, the only way to reset it was effectively to take it apart and reconstruct it in a later date configuration which you had worked out separately.

What if conscious states were like a much more complex version of this? What if you could only get to the state of thinking about tea-time through an appropriate series of earlier states? It might be that our whole conscious life is made up of a kind of rope of these threads of relevant states, stretching all the way back to the inscrutable autopoietic event in which our consciousness appeared out of nothing. If that were so, it might account for our sense of being responsible for our own actions: while the causes acting on inanimate objects are simply those that happen to be around in the environment at the time, a dominant factor in our own behaviour would be the self-contained stream of causality running along in our heads.

Moreover, an artificial intelligence would indeed have to start life in the same kind of unready and undefined state as a new baby and generate itself as it went along. It would also follow that a computer was a uniquely unsuitable machine for supporting such an entity. In principle a computer can go directly into any state: if you want to introduce a rule that state B follows state A, you have to do it through the program: so although a computer might be capable of exhibiting the right sequence of states which occur in thinking about tea (supposing those could be defined), the causal relationships between those states would be actually indirect. All you would get is a simulation, analogous to the simulation of motion provided by the rapid sequence of frames in a film.

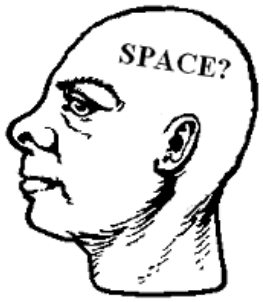
What about sleep? It seems a pretty good piece of evidence that human beings can indeed switch off and then resume when the right time comes. It might be hard to imagine coming into existence already thinking about a cup of tea; but awakening and starting at once to think about it is a thoroughly ordinary experience. It may be that some kind of mental activity continues even in sleep (and a theory of continuity might provide a new rationale for dreams); but people also come out of a coma, or dreamless unconsciousness. Unless we want to say that these are new people who merely inhabit the bodies and memories of their predecessors, it seems there is a difficulty.

Perhaps, in response, we could argue that beliefs persist even in sleep and coma. I may not think about anything while unconscious, but in some sense I go on believing that the Earth goes round the Sun, and not vice versa. Worryingly, in a similar sense beliefs continue even after death: does Luther still believe in God (discounting, for the sake of argument, the possibility of his surviving in a better place)? It seems odd to say so, but he certainly hasn't become an atheist in the last few centuries. Personally, however, I don't much like that line of argument, which seems to make our continuity both too absolute and too abstract at the same time. I would rather say that our continuity is not essentially disturbed if some of the states in the sequence persist, recorded in our memories and otherwise, through periods of inactivity.

Ultimately, though, the beginnings of consciousness remain as frustratingly unclear as most of its other aspects.

Non-Spatial Consciousness

4 June 2006



Colin McGinn is well-known as a leading champion of 'Mysterianism' - the view that consciousness is ultimately beyond our capacity to understand. On his version, human beings suffer from 'cognitive closure' in respect of consciousness: although there is really nothing unnatural or magic about the connection between consciousness and physics, our mental processes are just not up to the job of explaining it.

Besides this main case, McGinn has also offered another argument which buttresses his position, about the spatial, or rather the non-spatial nature of consciousness. In a nutshell, he suggests that while all physical events have both a time and a place, conscious events have only a time. It's hard for us to understand how this could be the case at all, let alone how, if it were truly the case, mental and physical events could be intimately related. Yet to preserve a monist account of the world, we need to suppose that mental events actually arise out of physical ones, or the other way round. McGinn, of course, does not really need to attempt to construct a way out of the impasse - he merely has to suggest that, as a matter of fact, it's incomprehensible to us.

There are some hints in the paper about what kind of theory of consciousness McGinn might be tempted to endorse if he didn't think theories of consciousness were beyond us. It's hard to see how mental stuff can have emerged from the non-mental world, he says: there's a kind of analogy with the Big Bang, wherein space emerged from non-space. It seems absurd to McGinn to suppose that the Big Bang was actually the beginning of time itself (in a footnote he rebukes those who have suggested this, for drawing a metaphysical conclusion from an epistemic premise, ie arguing that things we can't know about, such as times before the Big Bang, therefore don't exist); as a heady speculation, we might be tempted to guess that that same unknown stuff which supports our mental life also preceded and supported the emergence of the physical world. In his heart of hearts, it seems McGinn might like to believe that this esoteric ur-mental stuff of which we know nothing is, in ways we don't understand, the ultimate substrate of reality. However that may be, it seems to McGinn that some sort of unknown stuff is out there somehow: neither dualism nor monism can really be made to work, and most likely it is because our understanding of the physical world is missing out something big and important: 'The brain must have aspects that are not represented in our current physical world-view.' Aspects, as it happens, that that same brain is incapable of grasping.

I find the initial claim that consciousness is located in time but not space, appealing enough. It is intuitively plausible that earlier mental states cause later ones directly, but what could we make of the idea of adjacent mental states? Of course, as McGinn acknowledges, we can give a physical location to consciousness by identifying the brain which seems to support it, or by pointing out the physical objects to which we consciously attend: we can even point out that our pains seem to have pretty specific locations. But none of these points really touches the claim that consciousness in and of itself is no more located in space than, say, the number three. Nor, says McGinn, can we assume that consciousness exists in a kind of space we're merely unused to thinking about, as we do with some of the entities of physics: entities we never directly experience, but believe in because they are part of the best available explanation of the data.

The physical properties of consciousness are mysterious in a different, and less tractable sense, than the physical properties of, say, electrons.

Sophie R Allen, in a recent JCS piece ('A Space Oddity', Volume 13 No.4) has launched a limited but two-pronged attack on this position. Her main objection is to the idea that time can be regarded as ontologically separate from space. She examines four different conceptions of the relationship between space and time: standard relativity, Michael Tooley's diluted relativity (he apparently adds a privileged frame of reference, which seems to me a flat contradiction of the main idea); the views of Quentin Smith, and the traditional view of time as absolute. This is not a very comprehensive sampling of philosophical views about time, and the last two don't even seem to support Allen's case - both allowing a separation of space and time. She might perhaps have mounted another, pseudo-Leibizian argument: time is all about change, and without physical extension, no change is possible, ergo you can't have time without space. As it is, much the strongest of her arguments comes from orthodox relativity, partly because it is orthodox and authoritative, and partly because it really does involve abolishing the separation of time and space.

Even on that ground, however, I don't think McGinn need be unduly troubled. He could say that relativity applies only to physical entities - whether it applies to mental ones is an open question. It may well apply to us as physical animals, but we don't consciously experience time as relativistic. In fact, the gap between modern physics and our intuitive idea of the world is explicitly part of McGinn's argument. He points out that the progress of physics has taken our scientific view of the world gradually further and further from the 'folk' view (it's not just relativity - even the principles put forward by Galileo and by Newton are in some respects counter-intuitive). Isn't it likely, he says, that we need one more big leap to an even less common-sensical idea of the world - one which would encompass consciousness, but one which we are, alas, unfitted to make?

Allen, in fairness, does not claim to demolish McGinn's position, only suggest that it is unconvincing. However, to make much headway she really needs to establish that the idea of things existing in time but not space is more or less incoherent. I don't think it is: what about Christmas, for example?

The second prong of the attack is directed towards restoring the analogy between mental events and the unseen entities of physics. McGinn points out that while elementary particles and the like may be mysterious in some respects, we do conceive of them as existing in space and entering into broadly the same kind of spatial relationships as macroscopic entities (personally I can't help thinking in billiard-ball terms): mental entities are utterly different: they don't collide or exclude each other from spaces, or form triangles, or anything like that.

Yes, but, says Allen: how do we know that electrons and the rest operate in space? Only because they appear to have effects on familiar spatial objects which we can observe directly: and as a matter of fact, the spatial properties of some of these microscopic physical entities are decidedly queer and unlike those of the objects we see around us. Why shouldn't similar observations apply to mental entities? We know about them through their physical effects (at least in the case of other people), so they seem to have at least a nodding acquaintance with the physical world: why shouldn't they exist in some space (perhaps slightly deviant) of their own?

Fair enough - but the whole of this argument is, of course, something of a side-show. Even if we reject the analogy with the unseen entities of physics, we may still adhere to the idea that consciousness exists in a special space of its own. It doesn't have to be ordinary space - it might be a superficially plausible idea, for example, that each consciousness exists in its own quasi-solipsistic one-dimensional space (though perhaps harder to say clearly exactly what that would entail).

Allen concludes that McGinn may be right in thinking that we need a further paradigm shift in order to understand consciousness, but that he is altogether too pessimistic about our ability to achieve it. It's hard not to agree: why shouldn't we be able to crack the problem eventually?

But McGinn has a further argument up his sleeve. Citing P.F.Strawson, he suggests that our whole cognitive apparatus is based on spatial concepts. Fundamental ideas such as identity are rooted in ideas about 'being in the same place'. When we come to consciousness, therefore, we cannot avoid thinking about it spatially: but because it is non-spatial that means we are doomed to incomprehension.

This is a seductive, if depressing argument: but I have the same reservations about it as I do about McGinn's general thesis. If we were subject to cognitive closure in respect of consciousness, we should certainly be unable to solve the problem, but it seems to me we should equally be unable to perceive the problem. If our thinking were irredeemably spatial, we would happily work out whatever aspects of consciousness were amenable to such thinking, and never notice the rest. If the non-spatial aspects had spatial consequences, we should perceive them as independent spatial phenomena. We might indeed be forced to come up with some new, apparently arbitrary physical laws or constants, but we shouldn't find anything profoundly inscrutable. Let's not give up yet.

Ignorance Is The Answer

10 July 2006

Daniel Stoljar's new book Ignorance and Imagination makes a case for slug-like ignorance as the solution to our problems with consciousness. The real reason we have trouble with reconciling conscious experience with brute physical reality, he says, is that we just don't know enough. It's not that we need to crack some subtle piece of metaphysics; nor is it true that our minds aren't up to the job, as Colin McGinn suggests; nor yet is it a matter of finding some new and exotic additions to physics: there are just some key facts we don't have. The problem is epistemic. This seems a little like the position of John Searle, who says that consciousness arises from some properties of the brain we don't understand yet, but which are likely to be perfectly accessible to further scientific research.

To dramatise his point, Stoljar tells the story of slug-like creatures living on a vast mosaic pavement. There are two kinds of mosaic pieces: triangles and segments of circles. The slugs have sensory apparatus which can detect triangles or complete circles, but not segments of circles. Faced with this set-up, the slug theorists divide into several schools of thought. Some say that although the pavement contains squares, hexagons, and many other shapes, there are two fundamental elements: triangles, and circles. All the other shapes can be analysed down into triangles: the circles are an entirely separate affair and just have to be accepted as a feature of the pavement in their own right. Others say that ultimately circles can be reduced to triangles just like all the other shapes, so there is only one basic kind of thing in the pavement. A more radical group says that indeed triangles are all there are, and moreover the impression that there are circles is a mere illusion. Stoljar exploits the analogy cleverly, introducing slug equivalents for several of the most celebrated arguments about consciousness. The point is that the slugs can never get it right, because they simply don't know that all the circles are made up of segments, nor that other shapes made out of segments appear in various places.

The analogy is amusing and enlightening, but I'm not sure it actually serves Stoljar's purpose perfectly: it seems to me that the dualist slugs are actually right. Their knowledge of one of the two basic elements of the pavement is faulty, but they are correct in thinking that there are two basic elements. But Stoljar's main argument does not suffer from the same weakness. He defines the basic issue as being what he calls the logical problem. There are three plausible propositions involved.

- There are experiential truths (ie true statements about experiences)
- If there are any experiential truths, every experiential truth is entailed by some nonexperiential truth.
- If there are experiential truths, not every experiential truth is entailed by some nonexperiential truth

You can choose to believe any two of these, but whichever you choose, they will contradict the other one. I don't know whether this really deserves to be called the logical problem, but it is a neat formulation of a central problem, and it allows Stoljar to embark on a well-organised survey of the considerations which make each of the propositions attractive. Few would

seriously want to deny the first. If you want to believe in physicalism, the doctrine that everything in some sense comes down to physics, you have to adopt the second proposition. Arguments for the third are a little more complex. Stoljar presents two: the conceivability argument and the knowledge argument. The first is about the conceivability of having the physics without the experience: arguments from the possibility of zombies, or of inverted spectra. The knowledge argument is exemplified by the story of Mary the colour scientist.

Stoljar's treatment is careful and comprehensive, and one way or another he touches on most of the main arguments connected with consciousness in making his case. In terms of meaty pieces of discussion, the book is certainly value for money: but not overweight – Stoljar's prose is mercifully clear and straightforward, and there were only a couple of moments when I found myself wishing for a quick cut to the chase.

In fact, there was at least one place where I wanted more detail. One of the weaker areas, in my view, is Stoljar's consideration of conceivability. It is altogether implausible to me that conceivability straightforwardly implies possibility: the former concerns what can happen in my head while the latter concerns what can happen in the world out there. I think these arguments are only plausible if a restricted sense of possibility is specified. The fact that I can conceive of breaking the bank at Monte Carlo does not by any means imply that that feat is possible for me: but my ability imagine it might just imply that it is logically possible. I thought this aspect was not sufficiently acknowledged. Moreover, where Stoljar quotes examples of others using arguments from conceivability, I wasn't altogether convinced.

This arises with the historical precedents he quotes. As one means of making his thesis plausible, he offers examples of cases where people have drawn false philosophical conclusions because they simply lacked some key facts. Descartes said (not in English, but still) that it was "not conceivable that a machine should give an appropriately meaningful answer" to questions. Stoljar reads the argument as proceeding from Descartes' inability to conceive of such a thing to its impossibility: but Descartes, he says, just didn't know that certain kinds of linguistic competence are perfectly realisable from mere physics (after all, we set out to be physicalists, and that implies that we ourselves are mere machines at the end of the day). I'm not entirely sure that amounts to mere factual ignorance, but more fundamentally I'm not convinced Descartes actually meant to offer that argument. I suspect he was merely saying that it was just obvious that a machine couldn't answer intelligently.

Of course, what we really want to know is: what are these facts we need to know in order to understand consciousness? Stoljar's chosen mission here is to give a pared-down, generalised version of the epistemic view, so he does not offer a definite answer on this. He claims that to do so would be paradoxical, since you can't tell people unknown truths (yes, but come off it, Stoljar, you could tell us the truths which were unknown up until just now!). By way of further enhancing the attractions of his thesis, he does expound one possible kind of answer, which he calls Russellian. This is that the vital things which we don't know concern the categorical properties of objects – the inner, thing-in-itself properties which physics does not bear on. Stoljar notes that this line of reasoning might lead towards panpsychism, if we suspected that the categorical properties were in fact experiential. Although this has an undoubted neatness, panpsychists will be disappointed to hear that Stoljar takes this panpsychist leaning as one of the counts against the Russellian speculation. Personally, I prefer not to believe the Russellian line, in spite of its appeal, because it seems unlikely to me that we could ever gain any useful knowledge of the supposed categorical properties.

In the end, I think Stoljar achieves victory through the modesty of his ambitions: he doesn't even aim to prove that his view is true, he merely wants us to accept it as a plausible possibility. I do: but at the same time the nature of the chasm between experience and physics still leads me to think something more radical is really required.

Bayesian Consciousness

1 August 2006

The Max Planck Institute has set up a new project called the Bayesian Approach to Cognitive Systems, or BACS . Claiming to be inspired by the way real biological brains work, the project aims to create new artificial intelligence systems which can deal with high levels of real-world complexity; navigating through cluttered environments, tracking the fleeting expressions on a human face, and so on. There's no lack of confidence here; the project's proponents say they expect it to have a substantial impact on the economy and on society at large as the old dream of effective autonomous robots finally comes true.

It seems remarkable that the work of an eighteenth-century minister should now become such hot stuff technologically. Thomas Bayes is a rather mysterious figure, who published no scientific papers during his lifetime, yet somehow managed to become a member of the Royal Society: he must, I suppose, have been a great talker, or letter-writer. The theorem on which his fame rests, published only after his death, provides a way of calculating conditional probabilities. It gives a rational way of arriving at a probability based purely on what we know, rather than treating probability as an objective feature of the world which we can only assess correctly when we have all the data. If I want to know the odds of my taking out a green sock from my drawer, for example, I would normally want to know what socks were in there to begin with, but Bayes allows me to quantify the probability of green just on the basis of the half-dozen socks I have already pulled out. Philosophically, Bayes is naturally associated with the subjectivist point of view, which says that all probability is really just a matter of our limited knowledge - though given the lack of documentary evidence we can really only guess what his philosophical views might have been.

Universally accepted to begin with, then somewhat neglected, his ideas have been taken up in practical ways in recent years, and are now used by Google, for example.

Why are they so useful in this context? I think one reason is that they might offer a way round the notorious frame problem. This is one of the big barriers to progress with AI: the need to keep updating all your knowledge about the world every time anything changes. The problem as originally conceived was particularly bad because in addition to noting what had changed, you had to generate 'no change' statements for all the other items of knowledge about the world in your database. Daniel Dennett and others have reinterpreted this into a much wider philosophical problem about coping with real world knowledge.

Is there a solution? Well, the main reason the frame problem is so bad is because systems which rely on classical logic, on propositional and predicate calculus, cannot tolerate contradictions. They require every proposition in the database to be labelled either true or false: when a new proposition comes up, you have to test out the implications for all the existing ones in order to avoid a contradiction arising. It's clearly a bad thing to have conflicting propositions in the database, but the problem is made much worse by the fact that in classical logic there is a valid inference from a contradiction to anything (one of the counterintuitive aspects of

classical logic): this means any contradiction means complete disaster, with the system authorised to draw any conclusions at all. If you could have a different system, one that tolerated contradictions without falling apart, the problem would be circumvented, and that is why McCarthy, even as he was describing the frame problem for the first time, foresaw that the solution might lie in non-monotonic logics; that is, ones which don't require that everything be either true or false.

Where can we find a good non-monotonic system? It's not too difficult to extend normal logic to include a third value, neither true nor false, which we might call uncertain, or neutral: many would say that this is a more realistic model of reality. The snag with a trivalent logical system is that it sacrifices one of the main tactics available, namely that of deducing the falsity of a proposition from the fact that it leads to a contradiction. If there's always a third possibility, contradictions are hard to derive, and as a result trivalent logics are much less powerful tools for drawing new conclusions than classical logic (which isn't exactly a powerhouse of new insights itself).

Step forward, Bayesian methods. Now we can allocate a whole range of values to our propositions, representing the likelihood or the level of credence we assign to each. We don't need to re-evaluate everything when a new piece of information comes up, because head-on contradictions no longer arise; and we're no longer dealing in formal deductions anyway - instead we can use each new piece of evidence to make a rational adjustment in values. We don't get an instant conclusion, but what may be better: a gradual refinement of our beliefs whose ratings get more accurate the longer we go on. We can start without much information at all, and still draw reasonable conclusions.

This sounds pretty good, but in addition Bayesian methods are well established in the field of neural networks and there's some reason to think that this might be one of the ways real human brains work, especially in the case of perception. Rather than performing computations on visual data, it might well be that our brains use a Bayesian encoding, representing, say, the probability that we're seeing a straight edge at a certain distance from us, and using new data to update the relevant values.

This all seems excitingly plausible so far as basic cognitive functions are concerned, but what about consciousness itself? It seems to me quite a reasonable hypothesis that our opinions and beliefs are held in a Bayesian kind of way – mostly with varying degrees of certainty, and with a degree of tolerance for inconsistency. Changes in what we believe about the world do generally seem to arrive as the result of a relatively gradual accumulation of evidence, rather than through a sudden deductive insight.

But what about our phenomenal experience, the famous 'hard problem'? Here, as so often before, we seem to run aground a bit. I would have expected Bayesian perceptions to give us a rather cloudy, probabilistic view of the world, but instead we have a pretty clear and distinct one. What colour is the rose?

"Well, I'd say it looks 85% red, but it also looks 5% pink but in a poor light, and 4.5% orange. There's actually a distinct possibility of picture or model about the rose itself, and I can see an outside chance of hologram."

It very much isn't like that, or so it seems to me: our senses present us with a pretty unambiguous world: if there are mists, they are generally external. In all the cases of ambiguous

images I can think of, the brain enforces one interpretation at a time: you may be able to switch between seeing the shape as convex or concave, say, but you can't see it as evens either way.

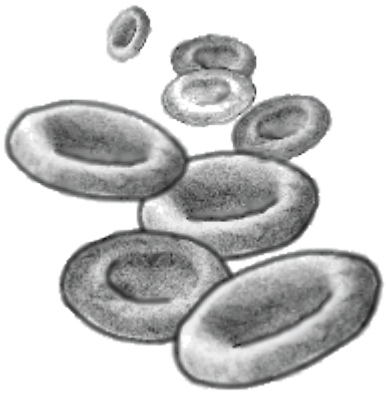
¶How much does that matter? We don't have to solve all the problems of cognitive science at once, and however confident it may be, I don't think the Max Planck Institute is attempting to. But I wonder why we are given this definite view of the world if the underlying mechanisms deal only in varying levels of probability? It can't be that way for no reason: we can only assume that there was some survival value in our brains working that way. It might be that this sharp focus is somehow a side-effect of full-blown consciousness, though I can't imagine why that should be; it might merely be that our perceptual systems are so good there's generally no point in confusing us with anything but the top probability.

But I think this might be a clue that for reasons which remain obscure, Bayesian methods alone will turn out to be not quite enough, even for some fairly basic examples of cognition. I won't be standing in the path of any of the early model robots, anyway.

I Feel It In My Blood

15 August 2006

'No,' said he, 'nor none
Of all these spirits, but myself alone,
Knows anything till he shall taste the blood.
But whomsoever you shall do that good,
He will the truth of all you wish unfold;
Who you envy it to will all withhold.'



Views of the afterlife seem to me to have become markedly more optimistic since ancient times; one of the most depressing versions must surely be the one in the Odyssey, where the dead prove to be mere shadows, devoid of all intelligence until they are provided with revivifying blood.

The idea of blood as the animating feature of the mind has acquired a modern echo of sorts in the theory put forward by Kenneth J. Dillon, namely that red blood cells provide the basic mechanism for a magnetoreceptive system that many animals possess to some degree: in human beings its clarity as a sense is somewhat clouded over, but it plays a vital role in the generation of consciousness.

I ought to admit to begin with that my knowledge of the theory is based purely on this extract from a longer and wide-ranging book, so I may well have missed some of the background: Dillon treats the existence of a magnetoreceptive sense in a wide variety of species, and of a human ability to perceive light with the skin, as established facts which require some explanation, whereas I wasn't aware that either was particularly well evidenced. However, the idea of thinking, or at least sensing, though the agency of the blood, has some appeal.

It's probably true that the role of blood in general has been unduly neglected in cognitive science: we know quite well that our emotions and patterns of thought are strongly influenced by hormones and other substances in the bloodstream, but with some honourable exceptions, the fact rarely gets much of a mention in discussions of consciousness. Magnetoreception is a slightly different matter, but after all, the view that phenomenal consciousness arises from an electromagnetic field, while not exactly mainstream, has been put forward by more than one author; if it is so, isn't it plausible that all those iron-rich red blood cells sweeping round our body have some role in the em field? The brain engrosses a disproportionate share of the body's blood supply - perhaps there's another reason for that, and it isn't just a matter of being a greedy energy consumer.

Dillon suggests that his theory might help to explain blindsight, the phenomenon in which people who are blind so far as conscious vision is concerned, can nevertheless 'guess' correctly the location of things they apparently can't see. Clearly if one had a misty magnetoreceptive sense, it might be able to fill in where your eyes failed you. But blindsight, so far as I know, only occurs in cases of certain kinds of specific, limited damage. A magnetoreceptive sense ought to work irrespective of the visual system, but there's no blindsight in cases where the eyes are destroyed, and I don't think it would work even for blindsight patients if they had their eyes covered or closed (though so far as I know, that variation on the experiment has not been carried out). Of course, one of the leading characteristics of blindsight is that it isn't conscious, so if the

red blood cells are supplying it that must presumably be distinct from any direct role they may play in consciousness.

Dillon also thinks that his theory might help with the binding problem, but I'm not convinced about this either. The binding problem is the issue of how the data from different senses gets combined into a smoothly running, uninterrupted and fully co-ordinated view of reality. The brain never gives us faulty lip-sync or sudden jumps and pauses in our view of events: but exactly how it pulls off this feat is unknown (some would argue that it doesn't have to, because the problem is misconceived). Now people often suggest, and I suspect this is what Dillon has in mind, various means by which the impressions from different senses could be brought together: but just bringing them together isn't enough: we need a method of working out what noises and what sights should be associated with the same instants, and of patching them together on the fly in a smooth sequence (and doing it all more or less instantaneously, because a lag of any appreciable size between reality and our view of it would clearly have a significant negative survival value!)

I also see some prima facie problems with the red blood cell theory. If the theory is true, oughtn't we to be much more aware of magnetic fields in our environment? So far as I know the human body is largely unresponsive to magnets. If we have any kind of magnetoreceptive apparatus, wouldn't we be sensitive in some way to electric motors, cathode ray tubes, and anything else with a significant magnetic component? Goodness knows what would happen to people undergoing an FMRI scan, since they are likely to be exposed to really whopping magnetic fields for an hour or more. But that point cuts both ways because, of course, FMRI would not be possible if blood didn't have significant magnetic properties which other tissues lack.

I also wonder why a magnetoreceptive facility would have remained so much in the background? You would think it would be a rather useful ability for an animal to have. It seems particularly strange that human beings should have it, or have had it, but allow it to remain almost completely dormant. Dillon suggests a number of possible reasons. We might need to be in a more trance-like state for it to work (an odd situation - we need to be less conscious in order to use the system which supports consciousness. Perhaps the role of supporting human-style consciousness is a new adaptation which has spoiled the sensory function.) Or our upright posture, habit of wearing clothes, or tendency to surround ourselves with artificial magnetic fields (ah, so it does have some effect!) has messed things up.

I'm no biologist, but I suspect that magnetoreceptive sensing isn't common because it doesn't work as well as some of the more usual senses. The creatures that indisputably have an electrical or magnetic sensory apparatus tend (I think) to need it because they live in murky conditions where vision is no good, or possibly because they want it to sense the earth's magnetic field as an aid to navigation. Human beings, living in well-lit terrestrial conditions, and not being migratory, wouldn't actually have much use for such a sense. In fact, I remain sceptical about their ever having had it to any noticeable degree.

I'm not convinced then - but you have to give Dillon credit for boldness and originality, qualities we're surely going to need before the mystery of consciousness is dispelled.'

Seeing Red

5 September 2006

Nicholas Humphrey recently produced a postscript to his book "Seeing Red". He remarks with mild regret that some readers perhaps will still not "get it": but if his theory is true, that's only to be expected. If he's right, the mysterious nature of qualia is really the point; if they were easy to understand, they wouldn't do their job properly – a tantalising suggestion.

Actually I suspect that the most troublesome point in Humphrey's theory, the one which most people tend to baulk at and the bit that's hardest to "get" is his view of sensation, which he distinguishes from mere perception (the former involves having a subjective experience, the latter is just a matter of affectless acquisition of facts). The problem is that Humphrey insists that sensation "has something of the character of a bodily action": it isn't just passive reception. When we see red, the sensation arises from us "redding", or to take a slightly more persuasive example, when we feel a pain, it's because we're hurting.

I must say that in the case of vision this seems completely counter-intuitive. Seeing colour feels very much like passive reception to me, but Humphries is very clear about its active character, referring to it as an example of agency. Agency surely requires control, the ability to act or refrain from acting; yet colour vision does not appear to be voluntary, like other examples of agency - I can't switch to monochrome vision in the same way I can decide to shut my eyes. It would be a shame if this put people off the rest of the theory, however, because it seems to me that the argument works just as well if we take "redding" to be a purely reactive affair, devoid of real agency.

What is "redding" and other similar actions, anyway? Although Humphrey doesn't describe it this way, we could regard it as a kind of second-order perception: a special sort of response to our own immediate reaction to stimuli. Once upon a time, our primitive single-celled ancestors might have responded to red light with a particular twitch of the membrane: as their descendants became more complex, Humphrey suggests, the initial reaction might have been disconnected from the twitch and used instead as the input to slightly more developed decision-making processes. Using your own reaction to stimuli as a source of information about the world has its limits, however, and ultimately our somewhat nearer ancestors would develop proper sensory apparatus separately. They would then have two channels providing information about the world: one a straightforward channel of perception, the other an indirect source based on monitoring the descendant of the old twitch-for-red reaction, still firing away somewhere in our brain. It's this internal monitoring of the phantom twitch which provides the qualia; but in fact it does more than that.

Humphrey points out that although strictly speaking the present moment is an instant of zero duration, we do not experience it that way. Our experience is of a short stretch of time, with recent events still to some degree in our thoughts. He refers to this as the "thick present", and suggests that a likely mechanism is a feedback circuit which causes present impressions to go on reverberating in our minds for a short time until they die away. It's a plausible idea, though I

suspect it may be a bit more complex than that: memory must surely be involved, and I would imagine that the level of attention we devote to things strongly affects the extent of their persistence and perhaps even the length of the perceived "thick present". Be that as it may, the real point is that Humphrey thinks his "redding" mechanism provides just the right kind of feedback loop to cause the kind of reverberations he is proposing.

So far so good, but why should monitoring our own vestigial twitches be anything like subjective experience? Why isn't it just like monitoring our own vestigial twitches? The fail-safe objection to any theory of qualia is the one that goes "Yes, but I can imagine all that stuff you describe happening in my brain or wherever, and me still not having any qualia". For once, though, there is at least a tentative answer.

Humphrey arrives at this answer by asking himself a different question. Why have we retained this complicated feedback mechanism, he asks. Evolution tends to weed out redundant devices, so there must be a strong presumption that a qualia-generating facility has some very definite survival value. Perhaps we can spot what this is by considering the case of blindsight, the strange phenomenon of people who can't see consciously, but can still point accurately to dots on a screen or pick up objects without fumbling. Humphrey has a long acquaintance with blindsight: in fact he described a case in an experimental monkey before the human version had been documented. He tells the poignant story of a woman whose sight defect was corrected late in life. Because her brain had had no visual inputs it had not developed the processing arrangements needed to deal with them, so although her eyes now worked perfectly she remained functionally blind. Or so it seemed: Humphrey suspected, and then confirmed by experiment, that although she had no conscious experience of vision, she did have a kind of general blindsight which, if she chose to use it, allowed her to perform all kinds of practical tasks which blindness had previously rendered impossible.

The strange thing is, she seemed unimpressed and unengaged by this new ability: did not enjoy using it and eventually chose to give up on it and revert to being the blind person which, in her subjective experience, she had always been. Why? Qualia, Humphrey suggests, are profoundly engaging: they make things seem to matter. Without them, his subject was unable to get herself to take any real interest in her blindsighted abilities; and without them we should all have an anaemic grey existence composed only of dull information. It seems to follow that philosophical zombies, people just like us but without qualia, would turn out to have a zombie-like lack of interest in life too, and hence be indifferent competitors in the battle for survival.

But why do qualia make things seem to matter? Because, in the final analysis, they are part of us; whereas our objective senses tell us about the external world, qualia are really our own internal reactions: surely it is only logical that we should find them more vivid and more important than simple facts about the world. In fact, they are also the basis of our sense of self, and help explain our devotion to the survival of that self, too.

We can now see why Humphrey half-expects people not to "get it". If it were clear to us that qualia were merely mental reactions, their valuable motivational effect would be dissipated, as would the useful delusion that we have a remarkable, immaterial self. It is only to be expected that evolution will have made the truth difficult to grasp, in our own interests.

This is a clever, cogent and rather ingenious argument, but I think I can see some rational reasons to doubt its truth. While perceiving qualia as part of ourselves might logically make them more important to us, I'm not sure that the idea of an immaterial self naturally encourages

us to take a keen interest in the survival of our bodies. In fact, we know quite well that the opposite is the case: people who believe strongly in the immortality of the soul are less concerned about death and may even seek out martyrdom.

More generally, does the wonderfulness of qualia really motivate us to be better survivors? It seems as likely to have us wasting time marvelling at the beauty of the daffodils while a tiger creeps up from behind: in survival terms it seems hard to beat the value of those grey but accurate facts. Actually, it is usually taken to be an essential property of qualia that they have no effect on our behaviour – if you're talking about something that changes the decisions we make, on this view, you're not talking about qualia at all.

I certainly have some reservations, in any case, about drawing many conclusions from blindsight examples. It seems quite likely to me that paying continuous attention to the subliminal or unconscious influences through which blindsight presumably works might be an exhausting, difficult, and uncertain business, one which the unfortunate woman in the case Humphrey quotes might easily have found burdensome, rather than simply unengaging through its lack of qualia.

There's something a little fishy, too, about the argument that qualia have a special impact because they are part of us. I can only care particularly in this way about things I know to be part of me – yet my redding has its special effect just because I don't recognise that it is internal (in fact, I imagine it to be a feature of red objects out there).

So although Humphrey's argument is a clever, interesting, and indeed quite persuasive one, I'm not quite convinced. Perhaps it's true that in one way or another evolution has destined me for disbelief in this case.

Tandem Triumphans

20 September 2006


Ted Honderich , as he promised a while ago, has returned in triumph to the periodical which woundingly rejected him a couple of years ago. Back then, Honderich's paper on 'Consciousness as Existence' failed the peer review for the Journal of Consciousness Studies: now, a whole issue is devoted to a target paper from him and eleven responses.

In the interval, Honderich has not been idle. He published "On Consciousness", a more substantial adumbration of the same theory, and others, notably Manzotti and Tonneau as mentioned here, have rallied to the banner of Radical Externalism which he has raised. There's really no such thing as final vindication for a philosopher, but this surely represents a remarkable improvement in the reception accorded to his views, and it must be highly gratifying: all the more so because, on the whole, the reception accorded his paper is a fairly friendly, positive one.


It's a good time for Honderich - he's on television in the UK this week presenting his views on the justification for terrorism. He is generally perceived as saying that it is justified, but of course it's more complicated than that - I think it's something along the lines of his cautiously asserting the possibility of a denial of the doctrine that there can in principle be no circumstance in which no less than one of the proffered justifications for terrorist acts can be seen on moral grounds to possess the qualifying characteristics mandated of it (whatever and in whichever theoretical context those may, adequately described no doubt in the relevant papers, be presumed to be) under those theories, not here to be elucidated but sufficiently indicated, perhaps, by such.

Perhaps that's a slightly unfair parody, but I find Honderich's expositions always suffer from a strange kind of fogginess. He seems to be asserting something very bold, but amongst all the meandering prose you get a certain impression of the fist not quite connecting. It was just the same, many years ago, with his well-received book, Punishment: it looked as if he thought all punishment was wrong, but nothing half so bold as that was ever quite asserted.

And now, consciousness: he seems to be saying, in fact he surely does say, that my consciousness of something just amounts to the thing existing, which is certainly bold: bold to the point of barminess - but it turns out that it actually amounts to the thing 'in a way' existing: three little words which actually stand for a hefty additional apparatus. Tim Crane, in his response, accuses Honderich of equivocating on this, and I find it hard not to agree: in some places Honderich seems to lay stress on the simplicity of the thesis that consciousness is existence, just existence: in others, he explains that what he has in mind is a new and puzzling mode of existence in a second or possibly a third metaphysical world with complicated relations of dependency on other worlds to sustain it.


I find Honderich's style rather engaging: it has a kind of gently self-deprecating humour. You're a bit unfair, aren't you? On the one hand you accuse him of being "foggy", and then you

complain that his brief statement of the main point is too terse: doesn't include all the details, and therefore conflicts or equivocates with what he says elsewhere.

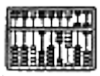
As Honderich fairly says, anyway, this particular piece isn't a detailed, definitive exposition, just a presentation of the theory: namely, that for you to be conscious of something is just for that thing to exist, not in the underlying physical world, nor in the form of firing neurons, but in another, perceptual world, dependent both on the physics and the neurons but separate from them. In other words, for you to see something, the actual atoms and molecules of it have to be there in the physical world, and your neurons have to be firing in whatever way they do when you notice something: but your consciousness does not consist in wither of those things: it consists in the object's being there in your perceptual world – not inside your skull.\r\nYou may find this a 'hefty apparatus', but I think it has a lot of intuitive appeal. Think of what people say when they talk about their perceptions – they use phrases like "there's something there", or "it's gone", even when they are not talking about an object but about the mere appearance itself.


Ah, mere appearances! Honderich can't actually cope with those, can he? Since he says that your awareness of something is equivalent to it's existing (in a way, blah de blah), it follows that illusions, dreams, imaginary things and mirages must all also exist (in a way, yacketty smacketty). Harold Brown illustrates the difficulty quite neatly by asking about Kanisza triangles (those ones which 'appear' when three 'pac-men' are positioned so as to define their corners but have no actual outlines). Do they exist (in a manner, yada yada)?

Honderich concedes that it isn't clear how these triangles fit into his conceptual framework, but he reckons it is a minor issue of the kind a friendly graduate student could sort out for him. Surely it's worse than that?

After all, one of the things Honderich particularly wants his theory to do is to banish sense-data style theories from the landscape. Many people have thought that we do not perceive the world directly, but through internal representations of some kind, and it seems to follow that all we ever really perceive is those representations. The main argument put forward for such views is the argument from error or illusion: if we perceived things directly, how could we ever be wrong about them? The error must creep in through our internal representation being different from reality. Honderich is very keen to deny this: indeed I think it is a major part of his motivation here. He asserts that perception is not perception of some representation or sense-data inside our head, it is the external existence of the thing perceived. But if you want to refute a theory whose main argument is based on problems over errors, your theory surely has to have a robust way of dealing with such errors?


I agree there's more to be done in clearing up issues like this, but Honderich has a number of options open to him. He denies that conscious awareness is the perception of internal representations, but that doesn't commit him to denying that there is any such thing as internal representations. It might well be perfectly reasonable to deny that the objects of perception are in the head while affirming that the objects of dreams and illusions are. This might imply that dreams and illusions are not perceptions, but what's wrong with that – I'd say they'd better not be! In responding to Paul Snowdon Honderich exploits a distinction between affective and perceptual consciousness, which seems a viable enough path to take.


OK. I suppose it's true that you can lash together a solution to any difficulty if you don't care about the additional overhead in complexity and ontological commitments that you incur. As a matter of fact, the basic ontological housekeeping of Radical Externalism is its weakest point if you ask me.

I mean, you've got this object of perception – let's be unimaginative and say it's a chair. It's not in the world of physics, it's not in my neurons, it's in this other place, this world of perception. Is that my world of perception, or the world of perception?

Let's assume to begin with that there's only one world of perception: is the chair I perceive there the same as the one you perceive? Strange if so, because it doesn't look quite the same to you as it does to me. These differences are not really errors, just differences in point of view and the like: so Honderich can't deploy whatever he may eventually come up with as a solution for errors (and if he could I think he'd find that his treatment of errors gradually eroded the rest of his theory away altogether). I think we're forced to conclude that the objects of our perception are different. This looks worryingly as if the chair itself might split into two, but Honderich can retain the identity of the single chair intact in the world of physics and just allow it to have, as it were, different avatars in the world of perception.\r\nHowever, if all the objects of my perception are separate from all the objects of yours, we might just as well say that we each have our own separate world of perception, containing the objects of perception special to us individually. So let's move on to the hypothesis that we each have our own world of perception with its own objects: our own are directly accessible to us, but not at all to other people (or we should all have a perfect kind of telepathy). Now you can call such worlds external if you like, but it means nothing: if they're particular to us and denied to everyone else, wouldn't it be more reasonable to describe them as internal?


I think the point is not about perceptions being specific to individuals, but about their being distinct from my personal neuronal activity. If we can agree that they're outside the skull, I think Honderich might not care all that much about your wanting to call them internal in some other vague sense. And your version of internality does strike me as pretty loose: my shoes are particular to me, but that doesn't make them internal. Don't get me wrong on all this – I'm not saying I'm signing up for membership of the Radical Externalists. But I would go along with what some of the respondents say: anything that cracks open the traditional ways we look at these things: gives us a new set of categories and concepts, must surely be welcome. Even if you don't like the externalism, you've surely got to give two cheers for the radicalism?'

The Narcissus Syndrome

30 September 2006

The Loebner prize, for the program best able to simulate a human being in conversation, along the lines of the celebrated Turing Test, has been won for the second year running by Rollo Carpenter's Jabberwacky, this time using 'Joan' a female personality. Last year's contest was mentioned here . Jabberwacky seems to be making steady progress: some media attention has been attracted recently by the visual people-simulations which have been commissioned to accompany the verbal output. George, last year's winning personality is now visible here.

My impression is that unless you deliberately set out to trap the program, it delivers quite long stretches of plausible, if rather evasive, dialogue. I suppose the evasiveness is an inevitable product of the computer not really understanding what you're talking about – it's like someone trying to bluff their way through a conversation about a book that they haven't actually read. It might disappear if the software were dealing with a limited domain, an area which it could 'know about' in more detail. That suggests that practical usefulness in some such restricted context (perhaps answering routine queries about a particular product, or being an interactive 'tour guide' in a museum) is no longer an unrealistic aspiration – though "really" passing the Turing Test in free conversation still is.

Should we be pleased or worried? Chatbot programs do not pretend to reproduce all the ineffable properties of real human thought and consciousness: they just aim to deliver good outputs by whatever means works. It has been suggested, however, that their resemblance to a full-fledged conscious being could have a subtly malign impact on our attitudes. The problem is that some people - many people, probably – are more than willing to attribute personhood to programs which haven't any claim to it. Daniel Dennett has remarked, in connection with his own erstwhile involvement with the Loebner, that the attitude of the interlocutor was often a more important factor than the quality of the chatbot. Some sceptical people devise cunning questions which rely on knowledge of the world, or the context of the dialogue, in ways which throw even the most sophisticated and well-prepared program: but many others accept almost any grammatical output as a human-like response. It seems there is something seductive about having our own image reflected in a machine; a kind of Narcissus effect which makes us fascinated with a dialogue in which we are really the only players.

This willingness to be deceived became evident with Joseph Weizenbaum's famous Eliza, the mother of all chatbots: he famously found his secretary conducting a long and deeply engaged conversation with the program, much to his horror. I saw the same effect on a smaller scale myself many years ago, in the days when an 80 by 25 display of glowing green characters was still the standard PC display technology. I had a simple program which drew a face on the screen using ascii characters: at random intervals it would raise an eyebrow, swivel its square eyes, smile, blink, and so on. Colleagues I had previously regarded as sensible took this thing to be far more sophisticated than it was: not human, obviously, but "maybe up to about the level of a tortoise" (it looked a bit chelonian).

We might well find, then, that if we start dealing with plausible chat-bots on a regular basis, we shall automatically start to think of them in much the same way as we think of human beings. But there are some obvious dangers in confusing people and simple machines. On the one hand, it might lead us to trust the advice of machines rather more than we should. We are already a bit prone to this, following the instructions of our in-car GPS navigation system even when it conflicts with common sense, and attributing a spurious authority to job evaluation or personality test programs which merely reflect back at us in a digested form the views we fed into them in the first place.

More seriously, we might find our instincts being tutored towards treating people the way we treat machines – as tools to be manipulated and used without any ethical significance. I think you could make a case that this sort of thing has already begun to happen: attitudes to euthanasia have certainly shifted a long way in recent times for example (if the thing's worn out, junk it), and utilitarian calculations about patients as generators of "quality of life" are far more overtly applied in medical contexts than would once have been the case. Moreover, it seems to be more readily accepted these days that moral responsibility is largely an illusion and that economic and social conditions determine the behaviour of criminals and heroes equally. Books and other works of art merely express the writer's place in the societal matrix rather than anything individual and ineffable. Does all this represent a welcome clarity about human nature, or a depressingly impoverished view of the world?

I understand the pessimistic view, which I think lies behind some people's distaste for the whole idea of AI. But in the end I think it under-rates the subtlety of people's attitudes. The fact is we are already used to a world in which all sorts of things which lack a human brain are treated in varying degrees as animate. Children do and don't believe that their toys have live personalities; the ancient Romans and many others believed in gods that were sort of people, and sort of mere embodiments of abstract qualities. In Japan, Shinto grants the status of Kami (god, spirit, ensouled thing) to all the most salient features of nature, without people becoming confused or losing their sense of human specialness. For that matter, humanoid robots have been a part of our culture for a long time now – the word itself is over eighty years old. Perhaps Weizenbaum's concern about his secretary was unnecessary: she may have regarded the friendly counsellor in the computer as no more real than the elusive Hearts players that help some Windows users pass the time these days. And… could it be possible that my old colleagues were to some extent just winding me up? I doubt if they would have gone back into a burning building to save tortoise-face, in any case. All in all, I think we can cope.

Of course, if we want to be really optimistic, there is another possibility – that dealing with chatbots every day will actually sharpen and improve our sense of the special qualities of real people…', 'The Narcissus Syndrome', '', 'publish', 'open', 'open', '', 'the-narcissus-syndrome', '', '',

Me do it

15 October 2006

Don't look at I: me is doing all the work. To hell with grammar: that's the point about consciousness, at least according to Tor Nørretranders. I must admit my heart sank just slightly when I first saw the title of his book *The User Illusion*. So many people want to denounce the self as an illusion these days! I think it's actually rather hard to deny that my self has some real substance – for most purposes selfhood seems to be an inoffensive, if not essential means of distinguishing between matters bearing on one human animal rather than all the others. Some of the sceptical arguments certainly have their appeal, but I generally feel that the best of them question the nature, rather than the existence, of the self.

Be that as it may, Nørretranders puts together a good case. It has a strong central theme, though it draws on arguments from several different sources. It's a wide-ranging book, in fact: in places it reminded me of Roger Penrose's tendency to go off on fascinating but slightly peripheral expositions. There's even a picture of a Turing machine, very similar to Penrose's, with the infinite tape heaped up in lines of boxes which disappear over the horizon (I worry slightly about the tangles that are liable to arise here – wouldn't it be better to have the tape hanging down into a bottomless void?)

The main perspective is of human beings as information processors, which means tackling the vexed question of what information really is and how it relates to reality. Nørretranders describes a conference where the participants boldly set out to get, as they put it "it from bit", a fantastically optimistic aspiration. If, as Frege found, we can't reduce maths to logic, how likely is it that we shall be able to reduce the actual existence of a particular apple to information?

Claude Shannon. of course, gave us a watertight way of doing hard calculations about information, but only by adopting a restricted definition, which relates it to order and entropy but excludes the whole idea of meaning, essential to the everyday conception of information. Quantifying information in this wider sense is formidably difficult. In ordinary human discourse, a few words can convey a tremendous quantity of information: in fact, in the right context a single symbol can speak volumes. An exclamation mark requires only a handful of bits, but when Victor Hugo and his publisher exchanged telegrams which read merely "?" and "!" a great deal was conveyed in both directions about the progress of Hugo's latest book.

To me, such examples show that there is something fundamentally wrong with the wider quantification project: Shannon, I suspect, probably did about all that can be done. If I have agreed on a code or convention with my contact, a single exclamation mark can mean, not just many different things, but absolutely anything whatever: does that mean it conveys an infinite amount of information? In fact, Victor Hugo could go on deducing further information from his publisher's message for an indefinite period: it told him, for example, that his publisher was still alive at the moment of composing the message: that he was alive a minute before that point, half a minute before that point, and... but you get the idea. The example may seem silly, but the problem is real.

Nørretranders takes a more optimistic view. Perhaps the real measure of information is not the content, but the work that had to be done to get that content; to whittle down the range of possible communications to the one we've chosen. In fact, and this is an idea which recurs throughout the rest of the book, perhaps the important thing is how much information we had to discard in arriving at our message. Nørretranders introduces the concept of exformation – all the extra contextual stuff which doesn't get included explicitly in the message, but which the recipient can infer or recognise without difficulty. This is strikingly like the Gricean implicatures which Sperber and Wilson have tried to quantify, but that other line of enquiry into the same territory does not get discussed here.

A second broad theme concerns the limitations of consciousness: we think our conscious minds are in control, but in fact there is plenty of evidence that the bit of us that does the talking and writing doesn't really do much else. Nørretranders quotes a wide range of evidence, mostly well-known to those of us who follow these things: blindsight, split brains, subliminal influences, the tendency of patients to produce a confabulated rationale for behaviour they weren't actually in control of, and Libet's finding that the brain is geared up to act before we have consciously decided on action. Nørretranders quotes a finding by Zimmerman that the maximum information flow of conscious sensory perception is a mere 40 bits per second, while the eye alone is sending ten million bits of information to the brain over the same period. All those millions of bits are surely there for some reason, but not, it seems, to support conscious decision making.

Nørretranders proposes a model in which the "me", the unconscious mind, is working away with all the reams of information provided by the senses: when it needs to communicate with another me, most of the information is discarded, and the tiny remnant transmitted between the two "I"s, or the two conscious minds, of the parties involved: only a tiny amount of bandwidth is necessary: but the recipient's "me" is able to recover the copious exformation which goes with the message. He suggests that the I is really similar in some respects to the "user illusion" which makes personal computers viable. We don't need, and couldn't cope, with knowing all the details of how our computer does what it does, how streams of digital numbers are moved around between specific locations in storage or various registers: instead we need a simple analogy with pieces of paper, wastebaskets, tools, and so on. The analogy can be pretty loose, or even downright misleading in certain respects, so long as it broadly conveys what is going on and reduces the formidable complexity of the actual processes to something we can manage.

I'm not sure that analogy is very precise. Generating an illusion of purposeful personhood seems a much more demanding business than making a complex computer process look like a simple real-life one. Who is the illusion for? Is it the unconscious me that is being fooled into thinking that there's a conscious I in charge? How do you delude someone who isn't conscious? Moreover, it's surely I that think I exist, not me, so the illusion must actually be for I's benefit: but if I'm illusory, how can I think anything, let alone fall victim to the illusion ofmy own existence?

\r\nIt's certainly true that a great many mental processes unfold unconsciously, but it's far from clear that these unconscious processes fit together to form a coherent unconscious me. Nørretranders may have done himself a disservice by quoting so many exotic examples: the split brain cases suggest the unconscious me must be confined to one hemisphere of the brain, but the other cases, such as the blindsight ones, seem to contradict that. Most surprising of all in this respect is the way Nørretranders invokes and apparently endorses Julian Jaynes' remarkable theory of the bicameral mind. According to Jaynes, we were actually not conscious until some point in ancient history: the part of our mind which we now perceive as our

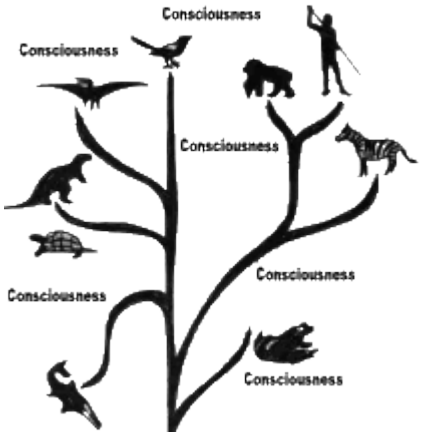
consciousness was previously interpreted as the voice of God or gods. But Jaynes has his preconscious people dealing with language readily: in fact the Aeneid is an example of the products of a bicameral mind (most implausibly, if you ask me, but still). That seems to fit very uneasily with an unconscious me which has run up the illusion of consciousness specifically in order to cope with language and other low-bandwith forms of communication.

It seems more likely that a whole series of unconscious processes are operating more or less separately. It also seems to me that the "I" as discussed by Nørretranders is drawn much too narrowly – it really amounts to no more than the part of us that does the talking. But I have had the experience, in stressful situations, of listening to my speaking part yak mindlessly on, while with my conscious mind I only wanted to shut up. When I play tennis, I don't think in words about where to hit the ball, or even very explicitly in other ways: but to say my tactics were unconscious would be going much too far (even for me). I suspect a more accurate model of how things work would be a complicated mess of unconscious processes feeding into a conscious level which is itself complex, with different levels which are clearly distinguishable, though still operating essentially as a unity.

That said, Nørretranders gives an insightful account, and covers an awful lot of ground. I don't know what me thought, but I enjoyed it.

A Spectrum of Consciousness

31 October 2006



There are many, many different ways of implementing consciousness, each with its own advantages and weaknesses, and it may well be that lots of them have been tried out during the course of evolution. So say Rodrick Wallace and Roger G. Wallace, in a paper full of daunting mathematics and airy speculations.

The Wallaces take Bernard J Baars' Global Workspace theory as their starting point. The Global Workspace theory sees consciousness as providing a general access function: a fleeting memory capacity that connects brain modules that otherwise function separately. It's as though consciousness were a noticeboard where the different functions can post problems, warnings, discoveries and data which other functions may be able to use or respond to. Consciousness is there to provide an overall integrative function.

The Wallaces have developed a detailed analysis of the requirements of such a function, and they have come to the conclusion that many different architectures and organisations are capable of giving rise to something with the essential properties. Indeed, they see no barrier to a slow form of consciousness in all sorts of creatures: a kind of 'paraconsciousness', as they call it. The details of this notion are tantalisingly vague, but it seems the idea is that even trees might have processes which allow them a very slow and very dim kind of awareness. The Wallaces also see analogues of the Global Workspace in systems made up of multiple organisms: societies or ant colonies. The idea that such higher-level systems have at least some of the properties of consciousness is reminiscent of Ned Block's 'Chinese Nation' thought experiment, in which the whole population of China is somehow dragooned into hand-simulating the nervous activity of a brain. The Chinese Nation idea, however, was supposed to demonstrate the weakness of functionalism by pointing out how absurd it is to think that this kind of gigantic hand-simulation of mental functions would give rise to anything like a mind. I suppose this perhaps shows how different our perceptions of intuitive plausibility can be.

¶The idea of higher level workspaces has a role, however, in overcoming some of the objections to the Wallace's ideas. The most obvious objection, of course, is that they take the Global Workspace for granted: they say it is rapidly becoming the most favoured model among researchers, but even if that is true, it's a long way short of being established. It remains quite possible that the theory is entirely wrong; or perhaps more likely, that an integrating function does exist, but does not constitute consciousness.

At the risk of being speciesist, it certainly looks as if the Wallaces are thinking of a form of consciousness somewhat short of the grand human version we are normally concerned with: they readily attribute it to a range of animals all the way down to cephalopods (who would surely achieve consciousness through a very different organisational structure from ours, given their tendency to rely on large neural ganglia distributed around the body rather than just a single central brain).

In fact the Wallaces are primarily interested in the palaeontology of consciousness, the road to which is, they say, wide and open (A rather optimistic conclusion, I think, since about the best

evidence we can hope for is the shape of a fossil skull. The Wallaces' own conclusions reinforce the common-sense assumption that you can't tell much about how a strange brain may have worked merely from its shape) . What strange forms of intellect may have flourished during the great Cambrian explosion, when many bizarre animals with strange body plans flourished!

However, they do have a view about the special qualities of purely human consciousness, and this is where the higher-level workspaces come in. It is likely that all the different forms of consciousness tried out by evolution had different strengths and weaknesses: most would have suffered from some areas of 'inattentional blindness', but the creatures that had the smallest blind spots in the least important places would have survived best. Perhaps the unique trick of human consciousness, the one which fitted it for a distinctly human way of surviving, was to have a special facility for tuning in to higher-level workspaces: for fitting into complex societies and networks of communication and decision making.

This idea is more or less an aside in the Wallaces' paper, but I can see some appeal in it, and I suspect it might in fact provide a more fruitful avenue to pursue than the attempt to reconstruct the minds of Acanthostega or Tulerpeton.

Does Consciousness Exist?

12 November 2006

There's a great collection of stuff about William James here, including the famous paper in which he introduced the concept of the 'stream of consciousness'. That idea has been more influential in literature than psychology, perhaps, but James' work enjoys tremendous respect among contemporary academics - much more than the works of some more recent thinkers whose ideas are tied to largely discredited theories.

Curiously, not all that many years after writing about the stream of consciousness, James produced a trenchant piece calling for the whole idea to be done away with. Consciousness, he wrote "is the name of a nonentity... a mere echo, the faint rumor left behind by the disappearing 'soul' upon the air of philosophy".

A century later, this still sounds a perplexingly radical stance. James himself rather wearily observed that his view would take a good deal of explanation to make it plausible, something he did not feel he had achieved in a single paper. I think that's pessimistic: the paper inevitably leaves a lot of metaphysical washing up to be done, but there's nothing half-baked about the central point. The best way of grasping what James means is to take his point in a historical context, as the quotation above suggests.

Once, straightforward dualism was the unquestioned orthodoxy: there were physical objects and spiritual objects and a great gulf lay between the two. Souls basically did the perceiving and the physical world did the being-perceived. Over time, however, the gulf began to close and the two sides began to get closer. But although the prevailing orthodoxy became increasingly monist, a distinction was always maintained, eventually boiling down to the difference between subject and object: and that difference is, in a word, consciousness. Subjects possess this mysterious substance, objects do not.

James calls for a further step towards a purer monism. There is no substantial difference between subject and object, nor (this is a little more difficult to accept) between thoughts and things. It is all a matter of functional relationships. He offers the analogy of paint: when it's in a tin, paint is clearly just a saleable commodity; when it's on canvas as part of a painting, it represents things, it has meaning, and all the rest. But it's still paint: it hasn't become spooky magic meaning paint; it hasn't been endowed with the miraculous stuff of subjectivity. It's still paint. The only difference is that on the canvas it now stands in certain relationships to certain things (objects represented, people) which it didn't while it was still in the tin.

Moreover, says James, so far as thoughts are concerned, aren't imaginary objects fundamentally similar to real ones? People have difficulty in accepting that a fictional rose is really red, or a hallucinatory knife really sharp, but why? James isn't arguing that imaginary objects are indistinguishable from real ones, but rather that they are distinguishable only through a difference in the relations between them: real knives cut real objects, while imaginary knives may or may not cut imaginary objects and don't cut real ones at all. Real knives, we notice, have stable and predictable relations with other real objects - that's what their reality amounts to, not a fundamental difference of substance.

Everything, it seems, reduces to experience, and in fact, if we reformulate our view of consciousness in those terms, it ceases to be problematic. So long as we recognise that it is a matter of relations within a monist world, we won't get into trouble: let's stop speaking of it as something in a world of its own.

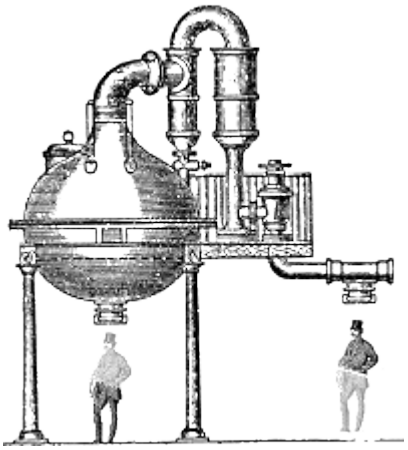
I think that once you get over the initial strangeness of this view, it seems pretty logical. But I see two problems. The first is that James speaks as though the different relations between things in his account of the world were more or less arbitrary: as it happens, this thing has the relations which make it an idea or a subject, and these things haven't. Surely it isn't quite like that. Some of these things have qualities which enable them to enter into the required relations, and others don't. A block of stone does not have the qualities needed to become a subject, still less a thought. On James' account, much of the difficulty and many of the challenges arising from the issues of consciousness transfer to the task of describing what these qualities are: and one may legitimately suspect that the dualistic magic he has banished from his account is likely to pop up again in one form or another when that task is undertaken.

Second, I think there are basic problems in saying that imaginary things can be red or sharp in the same sense as real ones. What imaginary knives have is the property of being thought-of-as-sharp: but that's clearly not the same as the property of being-sharp: real things can have the first in addition to, but distinct from, the second: in fact, they can have the property of being thought-of-as-sharp together with the property of not-really-being-sharp-at-all.

All the same, could it be true that we should do better to speak merely of experience, and stop talking about consciousness altogether?

Magic Scanners

22 November 2006



The New Scientist's special 50th birthday issue has a number of interesting pieces: Roger Penrose on the nature of reality, Patricia Churchland on free will, and others. It also includes a piece from Nick Bostrom, Director of the Future of Humanity Institute at Oxford, offering an argument to show that it is highly likely that we are all, in fact, living in a computer simulation. I must say this is not among the most convincing arguments I have ever read. It starts with the assumption that in due course we shall have enough computer power and programming skill to create simulations of our ancestors and their world, with the simulated ancestors having full consciousness and phenomenal experience. Just a matter of computer power and programming skill, you see. Now these sim ancestors, being just like their creators, will in due course set up a simulation of their own ancestors: and they in turn will set up another. In time, we shall have an indefinitely long sequence of simulations all nested within each other. At that point we ask ourselves how likely it is that we are living in the first step of this process: ie the unsimulated real world. Because there is only one such world, but an indefinitely large number of simulated ones, it proves to be almost certain that we are in one of the simulations.

Even if we accept that consciousness is fully computational (hardly uncontroversial); even if we agree that there are no important differences between reality and a simulation; isn't there something a little fishy about a simulation that contains a full copy of itself? How much of that computer power are we going to need? Could it be that to run an indefinitely large set of nested simulations, you would need an indefinitely large amount of capacity? I can't help wondering whether the future of humanity is in the best possible hands here.

Our favourite question, "What is Consciousness?" is addressed by Paul Broks in a kind of futuristic fantasy; our hero is about to have himself uploaded into a new youthful body, though he hasn't quite got the nerve to follow his daughter in getting the enhancements which would enable him to participate in the communal hive mind which is now available. I can't say I blame him about that - have you seen the general public? Would you honestly want to share brains with them?

The general drift of the piece is that really the self is more like a bundle of tumbleweed than anything with a fixed core (though to make the analogy precise I think balls of tumbleweed would have to be constantly shedding and picking up new material); as Broks pithily sums it up: I realised I was talking to myself but no-one was listening.

Broks quotes a hoary old thought experiment about scanners (yes, I'm getting to the point at last) - apparently it was originated by Parfit. Imagine, it goes, we have a scanner which can register all your physical details and then reconstitute you somewhere else, with every detail the same. Your life continues, you retain your memories; you barely notice that anything has happened, perhaps. This shows that there is no solid core to your identity: if the physics and the functional systems are moved elsewhere, you just go with them, and nothing is left behind.

It seems like bad manners to question the premises of a thought experiment, but can we take for granted the possibility of these Star Trek style scanners? I think we are entitled to at least an outline of how they might work: unless they are clearly possible in principle the argument doesn't get off the ground at all; yet I think there are some clear problems.

I think the idea is that the magic 'fluence of the scanner, when in reconstitution mode, causes the right kinds of stuff to condense out of nothing, already in the right places and with the right qualities. Perhaps I'm resurrecting a medieval prejudice against action at a distance, but I don't see how any conceivable apparatus could pull that off. There has to be some coherent causal chain which makes the right particles take up the right positions, which seem to imply that something has to intervene at those positions.

Perhaps that's the wrong way of looking at it: perhaps the scanner actually somehow shoots protons, neutrons and electrons into the right places layer by layer: perhaps it does indeed scan across a cross-section at a time and build the reconstituted person from the feet up. It would need some clever mechanism to take account of the fact that real people aren't made with a neatly laminated structure, and in fact it is going to have to be unbelievably accurate: given the importance of minute differences in the structure of neurons and the disposition of certain molecules within them, I suspect the tolerable error is actually zero. More awkwardly, it will have to allow for the fact that complex structures often require a particular assembly or construction sequence - this part has to be put in this way before that one goes in that way. If we start by scanning, let's say, Captain Kirk's feet into existence, the blood is going to start leaking out before we've done his shins, his tendons and muscles will lose their tension, and in general the whole thing will start falling apart in our hands. Perhaps, then, people get treated to the grisly sight of Kirk's skeleton being constituted first, to hold all the other bits up: but the scanner would have to put in some tendons and connective tissue to hold the bones together: in fact, since we can't have working muscles until the circulation is going, we might need extra ligatures or whatever which don't feature in the finished Kirk but get removed before the job is completed. There's still going to be a problem with that damn blood: we need to put all the vessels in place first and then fill them, which is tricky; and managing the filling and starting of the heart without a major problem will be difficult too.

But that won't do at all, in any case, because we're supposed to be reconstituting a dynamic system in full flight. It's difficult enough to reconstitute a snooker table, but what we have to do is bring the table into existence as it was a moment after someone played a shot, with the balls already in motion in various directions. Kirk's troublesome blood has to be flowing and all the right neurons have to be in mid-fire. The penalty if we can't do that (and I don't think we can) is that we have to reconstitute him in a slightly different, stable starting state, so that when Kirk is reconstituted he is unconscious and has suffered some loss of recent memory: when he comes round he doesn't remember getting into the scanner or why he wanted to be scanned: in fact, the idea that this is the original Kirk, rather than a good copy, suddenly seems much less plausible

It may be that getting used to computers has made us more ready to assume that reality is programmable, and hence to accept this kind of thought experiment. But the general idea has deeper roots. Even before consciousness became a fashionable topic, there was a long philosophical history of debating the nature of personal identity. One of the ideas that invariably got thrown out early in such discussions (and no doubt still does) was that our identity was, in the end, a matter of simple physical identity: this body really is me and vice versa. Physical identity is surely not the full story - but I wonder if it has been under-rated.

No more scanner arguments, please - unless they come with drawings and a technical specification.

Idealist Consciousness

3 December 2006

Most of those who have views about consciousness seem to be monist materialists. Indeed, the essential problem of consciousness is often formulated as being how we reconcile it with materialist physics, without any expectation of anyone's asking why we should want to. An embattled minority still rally round the dualist flag, but that's more or less it so far as popular metaphysical options are concerned.

But popularity isn't everything, and there is of course another position with an impeccable philosophical pedigree in the shape of idealism, the view most famously propounded by Bishop Berkeley with the maxim 'to be is to be perceived'. According to Berkeley things only exist because they feature in some mind: we are saved from a capricious dream-world of our own only by the mind of God, which encompasses and sustains everything, guaranteeing the consistency of the world and keeping things going when no human being is perceiving them. Axel Randrup and Peter B Lloyd, in different ways, have taken this same path, concluding that the solution to the problem of consciousness is easier if we assume that the mental world is the real one and the physical world a construction or fiction arising out of it.

What could motivate such a stance? I suspect that both Randrup and Lloyd have other reasons than consciousness for their idealist views, but both think it solves problems in that area which would otherwise pose formidable difficulties. Randrup sees a contradiction in the normal materialist view: it requires us to think that all of our mental life arose out of the evolution of simple matter, yet at the same time insists that that matter is wholly independent and separate from the mental stuff of consciousness. Lloyd thinks that the 'hard problem' becomes easy on an idealist view: the problem is about reconciling our experiences with the physical world, but if there is no physical world then we're home and dry.

Although I can see the appeal of these views, I don't think either argument is really convincing. It's not strictly contradictory to say that the mental arises from the physical and then reflects the physical. There is, indeed, a tension involved in having both mental and physical stuffs in your account, but that can be resolved as easily by materialism as idealism, or, a little less easily, by a plausible dualist system.

I don't think the hard problem is really a matter of reconciling our experience with the physical world so much as accounting for our phenomenal experience in any way whatever. It's as though we were examining a distant building in foggy conditions: we can see an elaborate roof high up which appears to be floating in space: we can also see the lower part of some columns near the ground, but nothing else is visible. Unfortunately the roof seems to be a completely different size and shape from the building suggested by the columns, and not even aligned with it. Well, says Lloyd, it's OK because the columns you can see are really just a kind of illusion: actually there's nothing there at all. That solves the consistency problem, but what we really wanted to know was what's holding the roof up: and that problem is at least as bad as before.

However, idealism does have some things going for it. Randrup, rightly I think, rejects the argument that idealism leads automatically to solipsism. It's true that if all we have to go on is our own perceptions, we could achieve a nice bit of ontological parsimony by denying that anyone else's perceptions exist. But materialists are not immune to that argument, since even they would generally accept that our experiences are the only evidence we have about the nature of external reality. They are obliged to justify their belief in the outside world by quoting the remarkable consistency and saliency of some of the hypothetical real-world entities which we take to be behind our perceptions: but idealists can take a similar line to support their belief in other people and even physical entities, without committing themselves to the independent reality of all these entities.

This does leave the idealists needing an explanation for the consistency of our experiences other than their having a source in an independent real world. For Randrup, this is a matter of intersubjectivity, a sharing of perceptions by groups of minds: for Lloyd, a more orthodox Berkeleyan, it arises from the metamind of which we are all parts, and which we can and perhaps should, call God.

Randrup is keen to explain that idealism does not mean we have to abandon all our science and just begin treating the world as being more or less an arbitrary dream: our theories need some reinterpretation but they remain valid. The trouble is, I think, that once the real causality of the world has been relocated elsewhere, the scientific story seems to be redundant. If the world were being spun out the imagination of a communal mind or metamind, why would it bother with all the details of chemistry and sub-atomic physics (though some aspects of modern physics might well be taken to resemble an imagined story whose inconsistencies had gradually got out of control)? Why would it stage a long history of evolution, and why would it take so long over working out something which it had thought up for itself?

Moreover, where are we left with consciousness? Whether you like it or not, one of the great attractions of computationalism, and to a lesser extent other materialist explanations is that they seem to offer a reasonably plausible and quite detailed explanation of what is going on - a real reduction of mentality to something easier to cope with. If idealism is true, we more or less have to accept the mental life as an unanalysed given, which is rather unsatisfying. In fairness, I think both Randrup and Lloyd would have at least some things to say by way of explaining consciousness, while none of the materialist theories on offer is anywhere near complete and satisfactory. But if we have to assess which of the two accounts looks more fully developed and offers more in the way of partial explanations, I don't think there's much contest.

Consciousness Yet to Come

15 December 2006

John Stewart has written an interesting paper on the future of consciousness - where will evolution take it next? His discussion is strongly rooted in the Global Workspace theory, but even if you don't subscribe to that school of thought it makes good sense. The prime function of consciousness, he says, is to give us new and ultimately better adaptive responses. In his view, it is the Global Workspace that allows the human mind to bring together resources which were not previously linked, but which can be used together in new and effective ways. This process becomes especially powerful when, in human beings, it gives rise to declarative knowledge - explicit, conscious thought. It's this ability that makes human beings able to frame complex and long-term plans, a huge benefit so far as survival is concerned.

But the use of declarative reasoning, thus far, is limited: in particular, while humans use it to find means of achieving their goals, they don't make very effective use of it in determining their goals in the first place: moreover, where they do think explicitly about their long-term objectives, they are very vulnerable to interference from hedonic impulses. In simpler language, the goals which you have chosen for good reasons of principle tend to be undermined by the pursuit of short-term pleasures thrown up by the more primitive parts of the mind. Passion ends up dominating reason.

This has a very familiar ring to it: for thousands of years philosophers and religious leaders have been enjoining us to raise our minds above ephemeral pleasures and seek more lofty goals. Stewart is happy to make the connection with religious thought and suggests that certain religious and contemplative traditions have developed techniques which we could use to help us develop our consciousness to a new level.

Interesting stuff, but there are a few points which seem dubious. In the first place, what he seems to be talking about is self-improvement rather than evolution: nothing we learn during our lifetime is transmitted with our genes. I suppose it might be the case that if enlightened thinking became a skill we all had to learn for our own good, those who had genes which especially fitted them for it might acquire an advantage and so the gene pool might move in that direction.

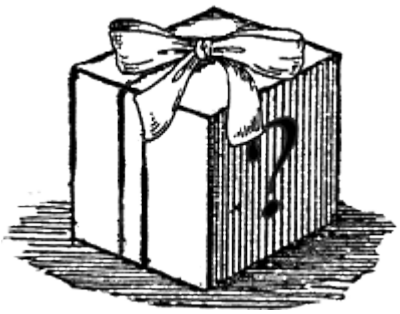
Second, does the kind of detached thinking he mentions really have survival value? I don't think Buddhist monks take up meditation in order to give their genes a better chance, and in fact some religious traditions have definite leanings towards celibacy and the sacrifice of one's own reproductive prospects. Stewart seems to think that if choose our goals rationally, we will choose ones that best serve our survival: but there's no purely rational reason to choose one goal rather than another - that's why we have built-in drives to begin with. It seems quite likely to me that a meditative, rational human race, operating on an unemotional level, might easily decide not to reproduce - or perhaps even not to eat. It may be that we need to stay stupid enough for our own good.

The question of where consciousness might go is an interesting one, though. What mystery gift might Mother Nature have in store for us? There is of course some debate about whether human

beings have escaped from evolutionary pressures for the foreseeable future, either by controlling their own environment, or perhaps just by moving around so much that there are none of the isolated populations which seem to play an important part in speciation.

Desiderata

20 December 2006



Since it's nearly Christmas, I propose to address the issue on an entirely frivolous level and present a Christmas wish-list of my top ten favourite improvements to consciousness...

Coming in at number 10 is Distributed Processing. Why can't we bud off a separate train of thought to deal with some subsidiary problem while we go on thinking about something else? We could have a whole tapestry of different threads separating and re-converging.

At number 9 I'd like an Indexed Memory. I don't need anything complex here - just a date-based system will do, so that I can remember what happened at any specified time and date.

Number 8 is a Vivid Memory. I'd like to be able to run personal flashbacks the way people in films do, where you relive the events being recalled in realistic detail.

At 7, I want Control of My Dreams. Not that I have terrible nightmares, you understand - I rarely recall any dreams at all. Some people apparently do have some control of some of their dreams, but I'd like to be able to script them completely. Purely for research purposes, of course. Well, partly for research purposes.

Number 6 is the ability to enter Altered States of Consciousness at will, without having to do any of that starving, meditating, or consumption of ayahuasca. If this is too much, I'd settle for a degree of control over my own endorphins.

At 5, I've got (or rather, I haven't yet got) Control of the Attention; the ability to concentrate on things, or indeed, ignore them.

Number 4 is Automation. There are many things I have to do laboriously in the forefront of my consciousness - calculations, writing routine letters, that sort of thing - which I wish I could delegate to unconscious functions which I'm sure are perfectly up to the job.

At 3, I want, ahem, Control of My Emotions. I don't actually want to be Mr Spock, but when someone treads on my foot I should like to be able stop being annoyed about it immediately instead of an hour and a half later.

Number 2 is Control of Unconscious Functions. I'd like to able to summon up placebo effects without being put through a confidence trick first; to stop the body's exaggerated and unhelpful response to burns and some other injuries, and rev my own heart up a bit when I know it's going to be needed but my unconscious mind doesn't.

At number 1, (since this is an intellectually oriented site), is Transparency, or Being Able To See How It Works. You cannot observe yourself in the process of composing your own thoughts, no matter how quickly you look over your shoulder. But wouldn't it be good if you could?