Review Article

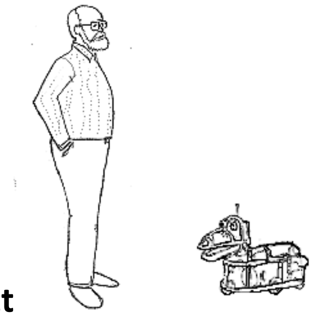
Eminent Entities: Short Accounts of Some Major Thinkers in Consciousness Studies

Peter Hankins*

ABSTRACT

I run a blog entitled "Conscious Entities" at http://consciousentities.com which is devoted to short discussions of some of the major thinkers and theories about consciousness. This is another small collection of my writings on consciousness which the editor of JCER very kindly selected to appear here. It contains my short accounts of six major thinkers in consciousness studies including Daniel Dennet, John Searle, David Chalmers, Colin McGinn, Roger Penrose & Gerald Edelman. In reading the books of these writers, I found I had views which were very clear, but also completely contradictory; so these pieces are written in the form of dialogues between a character I call Bitbucket (represented by the abacus) who is a hard-line materialist computational reductionist, and Blandula (the cherub) who leans towards dualism and mysterianism. (The last few words of each article, by the way, are actually quotes from the subject himself.)

Key Words: consciousness studies, people, Daniel Dennet, John Searle, David Chalmers, Colin McGinn, Roger Penrose, Gerald Edelman.



1. Daniel Dennett

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

Correspondence: Peter Hankins, Conscious Entities at http://consciousentities.com, London, UK.
E-mail: peter@consciousentities.com

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. Dennett'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 dispel 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 any more.

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…

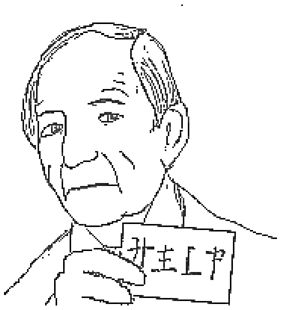
🧮Yes, very amusing I'm sure. I suppose I can sympathise with you to some extent. Grasping Dennett's ideas involves giving up a lot of cherished and ingrained notions, and I'm afraid you're just not ready (or perhaps able) to make the effort. But the suggestion that Dennett doesn't tell us what makes 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!



2. John Searle

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 clearto 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!

3. David Chalmers

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 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…

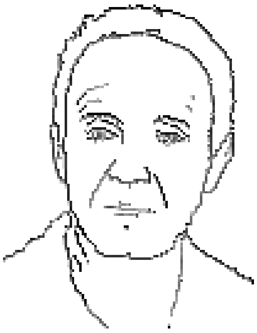
Chalmers does not actually endorse panpsychism, he just speculates about it. Do you think the idea is uninteresting ? Can you not accept that if philosophers aren't allowed to speculate, they're not going to achieve very much?

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…



4. Colin McGinn

Colin McGinn is probably 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.

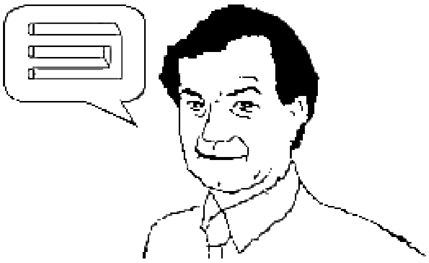
Any entity which makes a difference to the world must have some observable effects, and unless the Universe turns out to be deeply inexplicable in some way, these effects must follow some lawlike pattern. Once we've observed the effects and identified the pattern, we understand the entities as far as 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!



5. Roger Penrose

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.

1. 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'.
2. 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.
3. 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.

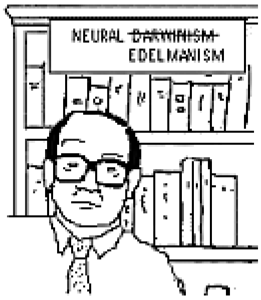
It isn’t quite as clear as that. There has been a tendency, right since the famous McCulloch and Pitts paper of 1947, to see neurons as simple switches, but the more we know about them the less plausible that seems. Actually there is some highly complex chemistry involved. Personally, I would also say 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…



6. Gerald Edelman

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 infiinite 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…