Conscious Entities

The Archive

2013 to 2015

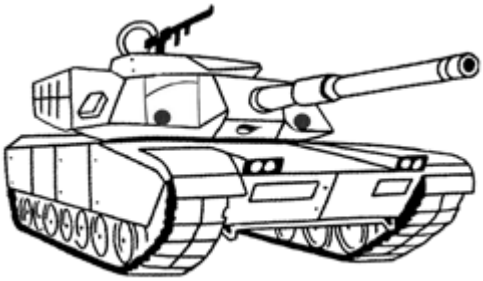
Welcome to Conscious Entities, the Archive

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 A few posts have been omitted, and quite a few are rather out of date, but I still enjoy reading them.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

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

If Guns Could Kill

6 January 2013



Back in November Human Rights Watch (HRW) published a report - Losing Humanity - which essentially called for a ban on killer robots – or more precisely on the development, production, and use of fully autonomous weapons, backing it up with a piece in the Washington Post. The argument was in essence that fully autonomous weapons are most probably not compatible with international conventions on responsible ethical military decision making, and that robots or machines lack (and perhaps always will lack) the qualities of emotional empathy and ethical judgement required to make decisions about human lives.

You might think that in certain respects that should be fairly uncontroversial. Even if you're optimistic about the future potential of robotic autonomy, the precautionary principle should dictate that we move with the greatest of caution when it comes to handing over lethal weapons . However, the New Yorker followed up with a piece which linked HRW's report with the emergence of driverless cars and argued that a ban was 'wildly unrealistic'. Instead, it said, we simply need to make machines ethical.

I found this quite annoying; not so much the suggestion as the idea that we are anywhere near being in a position to endow machines with ethical awareness. In the first place actual autonomy for robots is still a remote prospect (which I suppose ought to be comforting in a way). Machines that don't have a specified function and are left around to do whatever they decide is best, are not remotely viable at the moment, nor desirable. We don't let driverless cars argue with us about whether we should really go to the beach, and we don't let military machines decide to give up fighting and go into the lumber business.

Nor, for that matter, do we have a clear and uncontroversial theory of ethics of the kind we should need in order to simulate ethical awareness. So the New Yorker is proposing we start building something when we don't know how it works or even what it is with any clarity. The danger here, to my way of thinking, is that we might run up some simplistic gizmo and then convince ourselves we now have ethical machines, thereby by-passing the real dangers highlighted by HRW.

Funnily enough I agree with you that the proposal to endow machines with ethics is premature, but for completely different reasons. You think the project is impossible; I think it's irrelevant. Robots don't actually need the kind of ethics discussed here.

The New Yorker talks about cases where a driving robot might have to decide to sacrifice its own passengers to save a bus-load of orphans or something. That kind of thing never happens outside philosophers' thought experiments. In the real world you never know that you're inevitably going to kill either three bankers or twenty orphans - in every real driving situation you merely need to continue avoiding and minimising impact as much as you possibly can. The problems are practical, not ethical.

In the military sphere your intelligent missile robot isn't morally any different to a simpler one. People talk about autonomous weapons as though they are inherently dangerous. OK, a robot drone can go wrong and kill the wrong people, but so can a ballistic missile. There's never certainty about what you're going to hit. A WWII bomber had to go by the probability that most of its bombs would hit a proper target, not a bus full of orphans (although of course in the later stages of WWII they were targeting civilians too). Are the people who get killed by a conventional bomb that bounces the wrong way supposed to be comforted by the fact that they were killed by an accident rather than a mistaken decision? It's about probabilities, and we can get the probabilities of error by autonomous robots down to very low levels. In the long run intelligent autonomous weapons are going to be less likely to hit the wrong target than a missile simply lobbed in the general direction of the enemy.

Then we have the HRW's extraordinary claim that autonomous weapons are wrong because they lack emotions! They suggest that impulses of mercy and empathy, and unwillingness to shoot at one's own people sometimes intervene in human conflict, but could never do so if robots had the guns. This completely ignores the obvious fact that the emotions of hatred, fear, anger and greed are almost certainly what caused and sustain the conflict in the first place! Which soldier is more likely to behave ethically: one who is calm and rational, or one who is in the grip of strong emotions? Who will more probably observe the correct codes of military ethics, Mr Spock or a Viking berserker?

We know what war is good for (absolutely nothing). The costs of a war are always so high that a purely rational party would almost always choose not to fight. Even a bad bargain will nearly always be better than even a good war. We end up fighting for reasons that are emotional, and crucially because we know or fear that the enemy will react emotionally.

I think if you analyse the HRW statement enough it becomes clear that the real reason for wanting to ban autonomous weapons is simply fear; a sense that machines can't be trusted. There are two facets to this. The first and more reasonable is a fear that when machines fail, disaster may follow. A human being may hit the odd wrong target, but it goes no further: a little bug in some program might cause a robot to go on an endless killing spree. This is basically a fear of brittleness in machine behaviour, and there is a small amount of justification for it. It is true that some relatively unsophisticated linear programs rely on the assumptions built into their program and when those slip out of synch with reality things may go disastrously and unrecoverably wrong. But that's because they're bad programs, not a necessary feature of all autonomous systems and it is only cause for due caution and appropriate design and testing standards, not a ban.

The second facet, I suggest, is really a kind of primitive repugnance for the idea of a human's being killed by a lesser being; a secret sense that it is worse, somehow more grotesque, for twenty people to be killed by a thrashing robot than by a hysterical bank robber. Simply to describe this impulse is to show its absurdity.


It seems ethics are not important to robots be cause for you they're not important to anyone. But I'm pleased you agree that robots are outside the moral sphere.


Oh no, I don't say that. They don't currently need the kind of utilitarian calculus the New Yorker is on about, but I think it's inevitable that robots will eventually end up developing not one

but two separate codes of ethics. Neither of these will come from some sudden top-down philosophical insight – typical of you to propose that we suspend everything until the philosophy has been sorted out in a few thousand years or so - they'll be built up from rules of thumb and practical necessity.

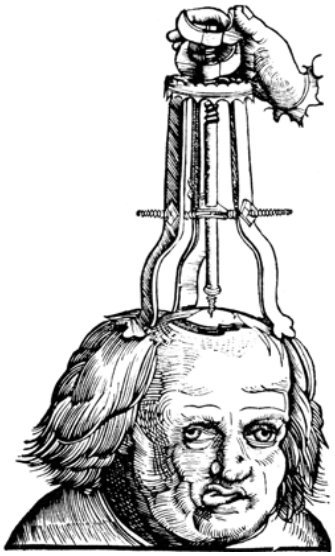
First, there'll be rules of best practice governing their interaction with humans. There may be some that will have to do with safety and the avoidance of brittleness and many, as Asimov foresaw, will essentially be about deferring to human beings. My guess is that they'll be in large part about remaining comprehensible to humans; there may be a duty to report , to provide rationales in terms that human beings can understand, and there may be a convention that when robots and humans work together, robots do things the human way, not using procedures too complex for the humans to follow, for example.

More interesting, when there's a real community of autonomous robots they are bound to evolve an ethics of their own. This is going to develop in the same sort of way as human ethics, but the conditions are going to be radically different. Human ethics were always dominated by the struggle for food and reproduction and the avoidance of death: those things won't matter as much in the robot system. But they will be happy dealing with very complex rules and a high level of game-theoretical understanding, whereas human beings have always tried to simplify things. They won't really be able to teach us their ethics; we may be able to deal with it intellectually but we'll never get it intuitively.

But for once, yes, I agree: we don't need to worry about that yet.'

Consciousness and Agony

20 January 2013



A piece in the Atlantic raises a question we have touched on before: anaesthesia. How do we know it works, how do we distinguish it from amnesia, and how much does it matter? What is supposed to happen is that when we get an anaesthetic injection, or gas, we simply stop feeling pain; possibly we stop feeling anything. It is certainly more complicated than that: the drugs used by doctors may simply stop us feeling pain - we more or less know with a local anaesthetic that that's sort of what happens - but they may also, or alternatively, stop us remembering, or caring about, the pain; they may also merely stop us moving or complaining. Some of the drugs used by anaesthetists apparently do only one of the three latter things, with no direct influence on the pain experience itself.

This matters most obviously when things go wrong, and in the middle of a surgical operation a patient resumes consciousness while remaining paralysed, experiencing all the terrible pain of being cut open without being able to give any sign of it. Because of the memory-erasing effects of some of the drugs used we cannot be sure how often this happens, but it is accepted that sometimes it does.

This is one of those areas where the apparently airy-fairy disputes of philosophy suddenly get real. A number of people are sceptical in principle about the reality of the self, of consciousness, and of subjective experience: but I think you need to be quite bold to stick to scepticism when it involves dismissing pain on the operating table as a mere conceptual confusion, something we need not be too concerned about.

There is, however, little agreement about some fundamental issues. Does it matter if I feel terrible pain, but then forget about it completely? I'd say yes, but a friend of mine takes the opposite view: if he doesn't remember it he considers that as good as not having had the experience in the first place. Of course I would have minded at the time, he says, but that's not what we're talking about; we're talking about whether I should mind now. Ex hypothesi, if I've forgotten, my mind is in exactly the same state as if I never had the pain, and it's incoherent to say I should worry about a non-existent difference.

I have to concede that my point of view opens up many more problems. Since memory is fallible, there might be lots of forgotten pains I should be worried about; but if I have a false memory of a pain that never happened, is that also a cause for concern? It's possible to be in a state of mind in which one feels a pain but somehow does not mind it: but what if I start minding about it later? What if I forget that I didn't mind? What if I did mind but the pain was actually illusory? What if I felt the pain unconsciously? And what if the memory then became conscious later? Or what if I felt it consciously but have only an unconscious memory? Do animals feel pain in the same way as we do? Do plants? If a drug dials my level of consciousness back to monkey level, dog level, chicken level, lizard level, ant level - does pain still matter? If I feel pain while operating on a protozoan mental level, does it matter more when I remember it on a fully human level? Do imaginative people suffer more (as, apparently, red-headed people tend to do)?

This may all sound stupidly speculative, but the questions are genuine and we're talking about whether I'm in agony or not. It would be great if we could bring all this out of philosophy zone and into science, but there are problems there too. The Atlantic article recounts difficulties with the BIS monitor, supposed to provide a simple numerical reading for the level of consciousness based on electroencephalograph readings.

Perhaps it's not surprising that the BIS has been questioned: an electroencephalograph is hardly cutting edge brain technology. A more fundamental problem is that it has no known theoretical basis: the algorithm is secret and the procedure is based entirely on empirical evidence with no underlying theory of conscious experience. Providing sound empirical evidence of subjective experience is obviously fraught with complications.

Step forward our old friend Giulio Tononi, whose theory of Phi, the measure of integrated information and hence, perhaps, of consciousness, is ideally suited to fill the conceptual gap. With Phi in one hand and modern scanning technology in the other, surely we can crack this one? I don't know whether the Phi theory is really up to the job: if it's true it might tell me why consciousness occurs in the brain and how much of it is going on, but it doesn't seem to explicate the actual nature of the pain experience, and that leaves us vulnerable because we're still reliant on the fallible reports and memories of the patients to establish our correlations. Could we ever be in a position where the patient complains of terrible pain and the doctor with the Tononi monitor tells them that actually he can prove they're not feeling any such thing?

Perhaps we can imagine a world where the doctor goes further. Great news, he says, we've established that you're a philosophical zombie: although you talk and behave as if you have feelings, you've actually never felt real pain at all! So no injection for you - we'll just strap you down and get on with it.

Feral Neurons

3 February 2013

Dan Dennett has confessed to a serious mistake, about homuncular functionalism.

An homunculus is literally a "little man". Some explanations of how the mind works include modules which are just assumed to be capable of carrying out the kind of functions which normally require the abilities of a complete human being. This is traditionally regarded as a fatal flaw equivalent to saying that something is done by "a little man in your head"; which is no use because it leaves us the job of explaining how the little man does it

Dennett, however, has defended homuncular explanations in certain circumstances. We can, he suggests, use a series of homunculi so long as they get gradually simpler with each step, and we end up with homunculi who are so simple we can see that they are only doing things a single neuron, or some other simple structure, might do.

That seems fair enough to me, except that I wouldn't call those little entities homunculi; they could better be called black boxes, perhaps. I think it is built into the concept of an homunculus that it has the full complement of human capacities. But that's sort of a quibble, and it could be that Dennett's defence of the little men has helped prevent people being scared away from legitimate "homuncular" hypotheses.

Anyway, he now says that he thinks he underestimated the neuron. He had been expecting that his chain or hierarchy of homunculi would end up with the kind of simple switch that a neuron was then widely taken to be; but he (or 'we', as he puts it) radically underestimated the complexity of neurons and their behaviour. He now thinks that they should be considered agents in their own right, competing for control and resources in a kind of pandemonium. This, of course, is not a radical departure for Dennett, harmonising nicely with his view of consciousness as a matter of 'multiple drafts'.

It has never been really clear to me how, in Dennett's theory, the struggle between multiple drafts ends up producing well-structured utterances, let alone a coherent personality, and the same problem is bound to arise with competing neurons. Dennett goes further and suggests, in what he presents as only the wildest of speculations, that human neurons might have some genetic switch turned on which re-enables some of the feral, selfish behaviour of their free-swimming cellular ancestors.

A resounding no to that, I think, for at least three reasons. First, it confuses their behaviour as cells, happily metabolising and growing, with their function as neurons, firing and transmitting across synapses. If neurons went feral it is the former that would go out of control, and as Dennett recognises, that's cancer rather than consciousness. Second, neurons are just too dependent to strike out on their own; they are surrounded, supported, and nurtured by a complex of glial cells which is often overlooked but which may well exert quite a detailed influence on neuronal firing. Neurons have neither the incentive nor the capacity to strike out on their own. Third, although the evolution of neurons is rather obscure, it seems probable that they are an opportunistic adaptation of cells originally specialised for detecting elusive

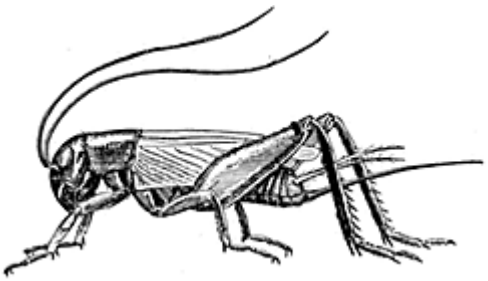
chemicals in the environment; so they may well be domesticated twice over, and not at all likely to retain any feral leanings. As I say, Dennett doesn't offer the idea very seriously, so I may be using a sledgehammer on butterflies.

Unfortunately Dennett repeats here a different error which I think he would do well to correct; the idea that the brain does massively parallel processing. This is only true, as I've said before, if by 'parallel processing' you mean something completely different to what it normally means in computing. Parallel processing in computers involves careful management of processes which are kept discrete, whereas the brain provides processes with complex and promiscuous linkages. The distinction between parallel and serial processing, moreover, just isn't that interesting at a deep theoretical level; parallel processing is just a handy technique for getting the same processes done a bit sooner; it's not something that could tell us anything about the nature of consciousness.

Always good to hear from Dennett, though. He says his next big project is about culture, probably involving memes. I'm not a big meme fan, but I look forward to it anyway.

Upon Thy Glimmering Thresholds

18 February 2013



I have been reading about the Brain Preservation Foundation (BPF), which hopes that chemical and other methods, including a refined version of plastination, will enable brains to be preserved with such fidelity that memories, personality, and even identity can be preserved.

This may well seem reminiscent of the older cryogenic preservation projects which have not always had a good press over recent years, though they still continue to operate and indeed have refined their processes somewhat. But although the BPF also has a vision of bringing people back to life after their natural death, it is in many ways a different kettle of fish. It does not itself offer any kind of service but merely seeks to promote research, and it does not expect to see a practical system for many years. In addition, it makes its case and addresses objections in a commendably clear and thoughtful way - see for example this blog post by John M Smart, co-founder of the BPF. Perhaps this is partly also to do with its impressive panel of advisors, which includes such names as Chalmers, Seung, and Eagleman, to mention only a few.

I have some reservations about the project, which fall into several categories; there are general concerns about the practicality of preservation, doubts about personal identity, and doubts about the claimed social value of letting people have a prolonged or renewed life; but there are positive factors, too.

There are clearly a lot of technical issues involved in preserving a brain (often said to be the most complex object in the universe) in all its detail, most of which I'm not competent to assess. I think the main general practical issue (if this counts as practical) is that although you might get a quite different impression from the popular press, we still don't really have a really clear idea of how the brain works - so in preserving it it's hard to know whether we're getting the right features. Clearly we would want the neuronal structure preserved in fine detail; but we keep finding out more about such matters as the incredibly complex sets of neurotransmitters that make the system work, about electrical interactions, and about the actual and possible role of astrocytes. If we're optimistic we may feel we're close to a working picture, but then we felt like that about genetics until the human genome was sequenced, and it's now becoming increasingly clear that we didn't know the half of it. Even without considering whether there might after all be something in the Penrose/Hameroff theory of unknown quantum mechanics operating in microtubules, or in similar ideas from outside the mainstream, there is a lot to think about. Of course the BPF can justly say that it is well aware of these issues , that they only reinforce the need for more research, and that working on preservation could well be a good way of pushing that research forward.

I think it's conceivable that there are also problems waiting to be discovered at a deeper level and that the brain can't even theoretically be 'frozen' in working shape - particularly if mental activity turns out to be inherently dynamic. This could happen in a couple of general ways. First, the brain could be like a zero-gravity box full of bouncing and colliding balls. You can't halt the activity in such a box and restart at an arbitrary point: you have to start at the point where the balls were thrown in. Second, the brain could be like the old astronomical clocks which were

geared together in such a way that they could not be reset; if they ever ran down and stopped, the only way to put them right again was to rebuild them. If either of these issues affects the brain, then it could not be restarted from its state at the last moment of consciousness, but only from some earlier state which might be a few minutes back, a few weeks, or in the worst case, the moment of birth! Now most of us would not mind losing just the last few minutes of our life if it meant we could then live on indefinitely, but even if that's all it amounts to, recreating that earlier viable state from the later one we had preserved might be extremely difficult - if that's the game we're in.

The good news on that is that all the empirical evidence suggests there is no such problem. People have come back from states in which brain activity had apparently run down very close to zero without any major long-term problem: so we can probably afford to be optimistic that a stopped brain is not a dead brain ipso facto.

More of a problem (perhaps) is the BPF's hope that preserving the brain might preserve personal identity. The philosophical literature on personal identity stretches back many centuries and I seem to remember that my own earlier self had some sophisticated views on it. As an undergraduate I think I developed a kind of morphic neo-nominalist position which dealt with nearly all the issues; but over the years I have become a caveman and my position now is more or less: see this? this rock – rock is rock – what your problem? To put it another way I think brute physical identity is essential to personal identity.

I'm aware, of course, that the atoms and molecules of our body are constantly changing – but so what? Why should we think reality resides at the most micro level? Those particles barely have identities themselves; they're more than half-way towards being mere mathematical constructions. People always say there's a good chance we're all breathing the odd molecule of oxygen that was once in the nostrils of Julius Caesar, but how would we know? Can you label a molecule? Can you recognise one? Can you even track where it goes? If I put three on a packing case, can you pick out the one you chose earlier? Isn't it part of the deal that two protons have identical attributes, apart from their spatial co-ordinates? They're not so much things as loci. Does talking about the same/different molecule, then, mean anything much? No, reality does not reside exclusively at the molecular level and I rest my case instead on the physical identity of certain neural structures, irrespective of their particle content. We are those critical neurons, I think.

Now you might think that the BPF is off to a flying start with me here, because instead of proposing to upload my mind onto magnetic media, they're aiming to preserve the echt physical neurons. But I do not think they are optimistic about the prospects of literally restarting the self-same set of neurons: rather they adopt a 'patternist' view in which it's the functional pattern of your mind that carries your identity. I doubt that: for one thing it seems to mean there could be as many of you as there are copies of the pattern. However, the strange thing is, I don't think the majority of people will actually care: rightly or wrongly they'll be just as happy with the prospect of a twin - really a kind of hyper-twin, far more like you than any real-life twin - as they would be with their own identity. Hey, they'll say, I'm not really the same person I used to be ten years ago anyway, any more than you are the same person as that callow young morphic ne0-nominalist. Perhaps when we are in the unprecedented situation of being able to copy ourselves our conception of identity must naturally change and loosen, though as we ontologists say, the idea certainly gives me the willies.

There's another deep problem in becoming immortal: I might run out. As it is, old people sometimes seem to repeat themselves or get stuck in a groove. Perhaps there comes a natural point when really you've said and thought everything important you're ever going to say and think, and any further lifespan is just going to be increasingly stale repetition. Perhaps the price of acquiring a really fresh lot of mental plasticity is, frankly, being a new person. I have once or twice met older people who, while fit and mentally agile, seemed to feel that the job of their life was basically complete, and while they didn't specially want to die, there wasn't really that much detaining them any more, either. It's a common observation, moreover, that old people sometimes seem to repeat themselves or get stuck in a groove.

I don't really know how far the idea of people 'running out' is true or how far failing memory and appetite for fresh exploration might simply be the product of waning vigour and physical energy, of a kind the BPF might hope to rectify (if rectification is the appropriate term). However, it does seem likely that even if it were true different people would run out at different stages, and probably few of us have completely exhausted our potential by the time we've done three score and ten. George Bernard Shaw took the view that three hundred years would be about the optimum lifespan, and it does feel like a comfortable benchmark: so even if there are natural limits to how far we should go on, it might be good to have the option of another couple of centuries.

That brings us on to the social benefits which the BPF suggests might accrue from having older minds around. They suggest that these older minds would be liberal and enlightened, and a force for progress, [Correction: John M. Smart has very courteously pointed out that the BPF doesn't suggest this at all. I don't quite know where I picked up the idea from, but I apologise for the error] which seems to fly in the face of many generations of experience, which is that old people tend to be increasingly conservative. Old scientists whose theories have been refuted don't usually give them up: they merely die in due course, still protesting that they were right, and make way for a new generation.

If one comes from a culture that reveres certain ancestors, the the prospect of being able, as it were, to bring back Benjamin Franklin to put the Supreme Court right on a couple of points about the Constitution might look pretty appealing; but how many subjects are the oldies going to be experts on? You may think now that you're a pretty hip grandpa for understanding Facebook: in fifty years, are you going to have any grasp at all of what's going on in a society mediated by electronic transactions on systems that haven't even been conceived of yet?

The good news here is that if the BPF is right, future generations will be able to take the old people out and put them away again as required like library books. Your future life may turn out to be a series of disconnected episodes several hundred years apart. You may find yourself now and then waking up in worlds where your senatorial views on certain matters are valued, but don't count on having the vote, if such a thing still exists in recognisable form.

All in all, it can't be bad to reward research, and it certainly can't be bad to think about the issues, which the BPF also does a good job on, so I wish them plenty of luck and success.

[Perhaps carelessly, I agreed to be a 'critical friend' to the BPF. This meant that for several years the top Google result for 'Peter Hankins' was my picture on the BPF website, and I had to repeatedly tell people that, no, I didn't really believe you could preserve brains and bring them back to life later.]

Mind-meld Rats

1 March 2013

A paper by Pais-Vieira, Lebedev, Kunicki, Wang, and Nicolelis has attracted a great deal of media attention. The BBC described it as 'literally mind-boggling'. It describes a series of experiments in which the minds of two rats were apparently melded to act as one.

Or does it? One rat, the 'encoder' was given a choice of levers to push – left or right (in some cases a more rat-friendly nose-activated switch was used instead of a lever). If it selected the correct one when cued, it got a reward in form of a few drops of water (it seems even lab rats are not getting the rewards they used to these days). Some of the rats learned to pick the right lever in 95% of cases and these went on to the next stage where the patterns of activation from their sensorimotor cortex as they pushed the right lever were picked up and transmitted.

Meanwhile 'decoder' rats had been fitted with similar brain implants and trained to respond to a series of impulses delivered in the same sensorimotor area by pressing the right lever. In this training stage they were not receiving impulses from another rat, just an artificially produced stream of blips. This phase of training apparently took about 45 days.

Finally, the two rats were joined up and lo: the impulses recorded from the 'encoder' rat, once delivered to the brain of the 'decoder' rat, enabled it to hit the right lever with up to 70% accuracy (you could get 50% from random pressing, of course, but it's still a significant improvement in performance). In one pointless variation, the encoder and decoder rats were in different labs thousands of miles apart; so what? Are we still amazed that electrical signals can be transmitted over long distances?

A couple of other aspects of the experiments seem odd to me. They did not have a control experiment where the signals went to a different part of the decoder rat's cortex, so we can't tell whether the use of the particular areas they settled on was significant. Second, they provided the encoder rat with incentives: it got water only when the decoder rat got it right. What was that meant to achieve, apart from making the encoder rat's life slightly worse than it already was? In essence, it encourages the encoder rat to develop effective signals: to step up the clarity and strength of the neural signals it was sending out. That may have helped to make the experiment a success, but it also detracts from any claim that what was being sent was normal neural activity.

So, what have we got, overall? Really, nothing to speak of. We're encouraged to think that the decoder rat was hearing the encoder's thoughts, or feeling its inclinations, or something of the kind, but there's clearly a much simpler explanation baked into the experiment: it was simply responding to electric impulses of a kind that it had already been trained to respond to (for 45 days, which must be the rat equivalent of post-doctorate levels of lever-pushing knowledge).

Given the lengthy training and selection of the rats, I don't think a 70% success rate is that amazing: it seems clear that they could have got a better rate if, instead of inserting precise neural connections, they had simply clipped an electrode to the decoder rat's left ear.

There's no evidence here of direct transmission of cognitive content: the simple information transferred is delivered via the association already trained into the 'decoder'. There's no decoding, and no communication in rat mentalese.

The discussion in the paper ends with the following remarkable proposition.¶

> *…in theory, channel accuracy can be increased if instead of a dyad a whole grid of multiple reciprocally interconnected brains are employed. Such a computing structure could define the first example of an organic computer capable of solving heuristic problems that would be deemed non-computable by a general Turing-machine. Future works will elucidate in detail the characteristics of this multi-brain system, its computational capabilities, and how it compares to other non-Turing computational architectures…*

Well, I'm boggling now.

CEMI Vindicated

18 March 2013

So it turns out consciousness really is an electric buzz around the brain.

JohnJoe McFadden claims his conscious electromagnetic field information (CEMI) theory - which says that consciousness lies in the brain's electromagnetic field - has now been borne out by a number of recent research findings. His paper is in the JCS, but a pdf can be accessed at Machines Like Us. We first discussed the CEMI theory eight years ago (can it really be that long?).

The case for CEMI is based in turn on the idea that synchronous neural firing can be shown to correlate with conscious awareness. Others have thought that lots of neurons firing in harmony at the right frequencies might be important of course; but the CEMI theory explains why it should be important by suggesting that synchronous firing produces effects in the endogenous magnetic field, which unsynchronised activity does not. If registering in that field is taken to be the same as presentation to consciousness we have a neat account of the phenomenon.

The first of the new studies quoted by McFadden, by Fujisawa et al, showed that fields of the kind generated by gamma oscillations in a slice of rat hippocampus affected the neuronal firing pattern. This demonstrates that neurons can influence each other significantly by electromagnetic means quite separate from 'normal' synaptic activity. The second, by Frohlich and McCormick, showed broadly similar influences of electromagnetic fields in the visual cortex of ferrets, supporting the claim that the endogenous fields provide a positive feedback loop that helps set up oscillatory networks. The third is research by Anastassiou et al which showed that neurons could influence each other through electric field effects: we discussed this 'ephaptic coupling' and pointed out its relevance to McFadden a couple of years ago (you read it here first, folks!).

So there's the evidence; what does it actually mean? I think McFadden is right to claim that the evidence for electromagnetic field effects on neuron firing is now too strong to ignore. At a minimum, it's something brain simulations will need to take into account. It's likely, moreover, that rather than simply being a nuisance factor, it actually plays some functional role in how networks of neurons are recruited and operate together. Anything more? McFadden suggests it may solve the binding problem; I'm not so sure. The binding problem is essentially the question of how information flows from different senses, processed at different speeds, with lags and gaps, somehow manage to end up in a smoothly coherent perception of reality with no jumps or lip-synch problems. Solving that problem may well involve bringing the activity of different neural assemblies together, but to me it's not clear how field effects could do anything other than smoosh all the inputs together, which is almost the opposite of what we want.

Speculatively McFadden suggests the EM field might be doing field computing, whatever that may be. He quotes a bizarre finding from the School of Cognitive & Computing Sciences (COGS) group at the University of Sussex. They used an evolutionary approach to develop a network which could perform a certain task, and then deleted the nodes which weren't playing any part in the the final network. Weirdly, they found that one of the essential nodes was not actually

connected to anything; yet removing it made the network stop working; put it back and the network worked again. They concluded that electromagnetic coupling must be playing a part. Of course electromagnetic effects in a field-programmable gate array (the equipment used by COGS) are not particularly likely to be anything like electromagnetic effects in the radically different physical substrate of neuronal tissue, but it does illustrate the general principle.

I still don't see that there are particularly good reasons to say that the EM field is the home of consciousness. For one thing consciousness is full of very complex content: while I can easily see how that complexity could be encoded in the fantastically complex patterns of neuron firing which go on in the cortex, it's harder to think that the EM field has a sufficiently elaborate structure. My consciousness is (in places) quite sharply defined and multi-layered, whereas a EM field seems more likely to provide a misty general glow. Perhaps the neurons provide the content and the EM field the subjectivity?

But one thing McFadden's theory cannot be is a solution to the 'Hard Problem' of subjective experience; his electromagnetic consciousness is playing a vital functional role in the operation of the brain, whereas qualia, strictly defined, have no causal effects. So much the worse for the theory of qualia, you might think; that just helps show that Dennett was right and the whole business of qualia is nonsensical. However, Sue Pockett, whose electromagnetic theory of consciousness is a kind of cousin of McFadden's, has jumped the other way on this, accepting that her own electromagnetic consciousness is epiphenomenal: it is produced by the brain but doesn't in turn produce any effects of its own; consciousness is a mere observer. This enables her to stay in the game so far as the Hard Problem is concerned, but of course it lands her with a different set of problems.

Perhaps in another eight years things will look very different - I rather hope so.

Bats Without Evolution

31 March 2013

Thomas Nagel is one of the panjandrums of consciousness, author of the classic paper 'What Is It Like To Be A Bat' and so a champion of qualia; but also an important figure in inspiring the Mysterian school of pessimism.

Now he has inspired new controversy with his book 'Mind and Cosmos: Why The Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False.' Probably the part of the book which has elicited the most negative reaction is the doubts Nagel expresses about evolution itself, or rather about the currently accepted view of it. It's not that Nagel disbelieves in evolution per se, but he thinks there are important gaps in its account; in particular he doesn't think it accounts satisfactorily for the origin of life, or for the availability of the large range of living forms on which natural selection has worked. He is not endorsing Intelligent Design but he thinks some of its proponents have arguments which deserve a wider and more sympathetic readership.

That does seem a bit alarming. It's true, I think, that we don't yet have a full and convincing story of how life came out of inert chemistry. I'd also agree that some of the theories put forward in the past - like the naked replicators championed by Richard Dawkins in the Selfish Gene - look a bit sketchy and optimistic. But none of that would stop me putting my money on the true story being a fully materialist one in which Darwinian evolution plays an early and crucial role. Nagel's second problem, with the way nature seems to have provided a remarkable fund of variation in organisms, of a kind which lent itself to the emergence of sophisticated organisms, just seems misconceived. He offers no statistical analysis or other reasoning as to why the standard account is unlikely, just mere incredulity. It seems amazing that that could have happened; well yes, it does, but then it also seems amazing that we're sitting on a huge oblate spheroid which is rotating and orbiting round an even vaster sphere of terrifying thermonuclear activity; but there it is.

The real problem is that Nagel wants these doubts (together with some more specific objections to standard materialism) to justify a large metaphysical change in our conception of the whole cosmos. His book professes only to offer tentative and inadequately imaginative speculations, and the discussion is largely at a meta-theoretical level - he isn't telling us what he thinks is the case, he's discussing the kinds of theory that could in principle be advocated - but it's clear enough what kind of theory he would prefer to reductive materialism. What he leans towards is a teleological theory; one in which some underlying principle drives the world towards a particular goal. He does not want this to be an intentional goal; he does not want God in the picture or any other Designer; rather he wants there to be a natural push towards value; value being conceived as a sort of goodness or moral utility, although this part of the speculative potential theory clearly needs development.

Nagel's critique of evolution may seem alarmingly misplaced, but the idea of introducing teleology to fill the gaps seems really astonishing. What, we're going to abandon the idea that DNA came together through natural selection and say instead that it came together because it sort of wanted to or was sort of meant to? The history of science has been a history of driving out teleological explanations - and the reason that represents progress is that teleological explanations are just not very good; they are usually vacuous and provide no real insight or predictive power.

In some ways what Nagel is after seems like an inverted law of entropy. Instead of things running down and tending to disorder, he wants there to be something built into the cosmos that shoves things towards elaboration, complexity, and indeed self-awareness (he positions the evolution of human consciousness as a peak of the process, and likens it to the Universe waking up). In itself that vision is quite appealing, but Nagel wants it to be driven by the worst kind of teleology.

I'm not sure what the official ontological status of the law of entropy is - it could be a meta-law which says the laws of nature must be such that entropy always increases, or it could be something that emerges from those laws (perhaps from any viable set of laws) - but it is definitely fully compatible with the rest of physics. If Nagel's new teleology worked like this, it might be viable, but he actually supposes it is going to have to discreetly intervene at some point and turn events in a direction other than the one mere physics would have dictated. This seems a disastrous requirement. If there's one thing the whole weight of science goes to prove, it's that the laws of physics are not intermittent or interruptable; every experiment ever conducted has contributed evidence that they are consistent and complete. Yes, there are some places in quantum physics or wherever where some might hope to smuggle in a bit of jiggery-pokery, but I think on examination even these recondite areas offer no real hope of a loophole.

This is a general issue with Nagel's case. We can sympathise with the view that evolution is not a Theory of Everything, but the other theories we need should be compatible with the broadly materialist world view which, despite some problems, is really the only fully-worked out one we've got: but Nagel hankers after something stranger and thinner.

What about those other theories? Nagel isn't basing his argument simply on his doubts about evolution; he has three places in which he thinks the standard materialist view is just not adequate. Consciousness, unsurprisingly, is one; cognitive thought, more unexpectedly, is another; and the third is his concept of value. Let's consider what he has to say about each.

On consciousness the basic argument is simply that our inner experience is just inaccessible to science. We still can't get inside the heads of those bats, and we can't really get inside anyone's except the one we have direct experience of - our own. Nagel briefly considers the history of the problem and the theory of psychophysical identity put forward by Place and Smart, but nothing in that line satisfies him, and I think it's clear that nothing of the kind could, because nothing can take away the option of saying "yes, that's all very well, but it doesn't cover this here, this current experience of mine". Interestingly, Nagel says he actually suspects the connection between mental and physical is not in fact contingent, but the result of a deep connection which unfortunately is obscured by our current conceptual framework; so given a revolution in that framework he seems to allow that psychophysical identity could after all be seen to be true. I'm surprised by that because it seems to me that Nagel is in a place beyond the reach of any conceptual rearrangement (that cuts both ways - Nagel can't be drawn out by argument, but equally if someone were simply to deny there is "something it is like" to see red, I don't think

Nagel would have anything further to say to them either); but perhaps we should feel very faintly encouraged.

At any rate, Nagel argues (and few will resist) that if consciousness is indeed physically inexplicable in this way the problem cannot be sealed off in the mind; it must creep out and infect our ideas about everything, because we have to give accounts of how consciousness evolved, and how it fits into our notions of physical reality. The answer to that in short is of course that as far as Nagel is concerned it can't be done, but getting there through his review of the possibilities is quite a ride.

Cognition is the second problem, and somewhat unexpected: that's supposed to be the Easy Problem, isn't it? Nagel draws a distinction between the simple kind of reactions which relate directly to survival and the more foresighted and detached general cognition which he sees as more or less limited to human beings. He doubts that the latter is a natural product of simple evolution, which sort of echoes the doubts of Wallace, the co-discoverer of Darwin's theory; in later life he took the view that evolution could not explain the human mind because cavemen simply didn't need to be that bright and would not have been under evolutionary pressure to spend energy on such massive intellectual capacity.

Nagel sees a distinction between faculties like sight, and that of reason. Our eyes present us with information about the world; we know it may be wrong now and then, but we're rationally able to trust our vision because we know how it works and we know that evolution has equipped us with visual systems that pick up things relevant to our survival.

Our reasoning powers are different. We need them in order to justify anything to ourselves; but we can't use them to validate themselves without circularity. It's no good saying our reason must be serviceable because otherwise evolution wouldn't have produced it, because we need to use our reason to get to belief in evolution. In short, our faith in our own cognitive powers must and does rely on something else, something of which a separate, non-evolutionary account must be given.

There's something odd about this line of argument. Do we really look to evolution to validate our abilities? I have a liver thanks to evolution, but its splendid functional abilities are explained in another realm, that of biochemistry. I don't think we trust our eyes because of evolution (people found reasons to believe their eyes before Darwin came along). So yes, our cognitive abilities do, on one level, need to be understood in terms of an explanatory realm separate from evolutionary theory - one that has to do with logic, induction, and other less formal processes. It's also true that we haven't yet got a full and agreed account of how all that works - although, you know, we have a few ideas.

But surely not even the most radical evolutionary theorists claim that the theory validates our powers of reasoning - it simply explains how we got them. If Nagel merely wants to remind us that the 'easy' problem still exists, well and good - but that's not much of a hit against materialism, still less evolution.

The third big problem is 'value'" a term which here confusingly covers three distinct things: first, the target for Nagel's teleological theory - the thing the cosmos hypothetically seeks to maximise; second, the general quality of the desiderata we all seek (food, shelter, sex, etc); third, the general object of ethics, somewhat in the sense that people talk about "our values". These three things may well be linked, but they are not, prima facie, identical. However Nagel wants to sweep them all up in a general concept of something loosely motivating which is

absent from the standard materialist account. He quotes with approval an argument by Sharon Street about moral realism, with the small proviso that he wants to reverse it.

Street's argument is complex, but the twice-summarised gist appears to be that 'value' as something with a real existence in a realm of its own is incompatible with evolution because evolution happens in the real material world and could not be affected by it. Street draws the conclusion that since evolution is true, moral realism in this sense is false, whereas Nagel concludes that since moral realism is just evidently true, evolution can't be quite right.

Myself I see no need to bring evolution into it. If moral value exists in a realm separate from material events, it can't affect our material behaviour, so we have an immediate radical problem already, long before we need to start worrying about such matters as the longer-term history of life on earth.¶I said that I think 'value' is actually three things, and I think we need three different answers. First, yes, we need an account of our drives and motivations. But I feel pretty confident that that can be delivered in a standard materialist framework; if we lay aside the special problems around conscious motivation I would even venture to say that I don't see a huge problem; we can already account pretty well for a lot of 'value' driven behaviour, from tropisms in plants up through reflexes and instincts, to at least an outline idea of quirte complex behaviours. Second, yes, we also need an account of moral agency; and I think Nagel is right to make a linkage with philosophy of mind and consciousness. This is a large subject in itself; it could be that morality turns out in the end to need a special realm of its own which gives rise to problems for materialism, but Nagel says nothing that persuades me that is so, and things look far more promising and less problematic in the opposite direction. Finally, we have the fuel for Nagel's teleology; not wanted at all, in my view: an unnecessary ontological commitment which buys us nothing we want in the way of insight or explanation.

To sum up; this has been a pretty negative account. I think Nagel consistently overstates the claims of evolution and so ends up fighting some straw men. He doesn't have a developed positive case to offer; what he does suggest is unattractive, and I must admit that I think in the end his negative arguments are mainly mistaken. He does articulate some of the remaining problems for materialism, and he does put some fresh points, which is a worthy achievement. I sympathise with his view that evolutionary arguments have at times been misapplied, and I admire his boldness in swimming against the tide. I do think the book is likely to become a landmark, a defining statement of the anti-materialist case. However, that case doesn't, in my opinion, come out of it looking very good, and by associating it so strongly with misplaced anti-evolutionary sentiment, Nagel may possibly have done it more harm than good.

Are Babies Conscious?

21 April 2013

Probably, according to a new paper by Sid Kouider et al. Babies can't report their own mental states, so they can't confirm it for us explicitly: and new-borns are regarded by the medical profession as unknowing bags of instinct and reflex, with fond mothers quite deluded in thinking their brand-new offspring recognises anything or smiles at them (it's just wind, or some random muscular grimace). However, Kouider and his associates tracked ERPs (event-related potentials) in the brains of infants at 5, 12, and 15 months and found responses similar to those of adults, albeit slower and weaker, especially in the younger babies. So although few babies are able to pull off the legendary feat of St Nicholas, who apparently uttered a perfectly-articulated prayer immediately on emerging from the womb, they are probably more aware than we may have thought.

Kouider has a bit of form on consciousness: last year Trends in Cognitive Sciences carried a dialogue between him and Ned Block. This arose out of a claim by Block that classic experiments by Sperling support the richness of phenomenal consciousness as compared with access consciousness. Block is probably best known for introducing the distinction between phenomenal, or p-consciousness, and access, or a-consciousness, into philosophy of mind. Roughly speaking, we can say that p-consciousness is Hard Problem consciousness, to do with our subjective experience, and a-consciousness is Easy Problem consciousness, the kind that plays a functional role in decision-making and so on.

Sperling, some fifty years ago, showed that subjects shown an array of letters could report only 3 or 4 of them; but when cued to think of a particular line, they were able to report 3 or 4 from that line. They must therefore have had some image of more of the array - probably the whole array - than 3 or 4 items but could only ever report that many.

Block's analysis is that the whole array was in phenomenal consciousness, but access consciousness could only ever get 3 to 4 items from it. This is apparently supported by what test subjects tended to say: they often claimed to be conscious of the whole array at the time but not able to recall more than a few of the items (although the Sperling experiments show they could report the quota of items afterwards from any row, indicating that the problem was not really with recall but with access).

Kouider, among others, rejected this view, suggesting instead that the full array is retained, not in phenomenal consciousness, but in unconscious storage. In his view it's not necessary to invoke phenomenal consciousness, which is an unverifiable addition which we're better off without. The subjects' feeling that they had been aware of the whole array can be attributed to a sort of illusion; you don't notice the absence of things you're not aware of, any more than you can see whether the refrigerator light goes off when the door is closed.

It's tempting to think that the dispute is at least aggravated by terminology; everyone agrees that information about the array is retained mentally in a place other than the active forefront of the mind; isn't the argument merely about whether we call that place phenomenal or un-conscious? That doesn't seem altogether satisfactory, though, if we take phenomenal consciousness seriously - we are talking about whether the subjects are right or wrong about

the contents of their own consciousness, which seems to be a matter of substance. I wonder though, whether there is some unhelpful reification going on - are phenomenal consciousness and the unconscious really two 'places'? Is it really more a matter of retaining information phenomenally or unconsciously? That might be a slightly more promising perspective, although I also think that mental states are generally a trackless swamp and a dispute with only two alternatives may actually be underselling the problem (could it be kept both unconsciously and phenomenally? Could it be not unconsciously but subconsciously? Could the route from unconscious storage to access consciousness lead via phenomenal consciousness?)

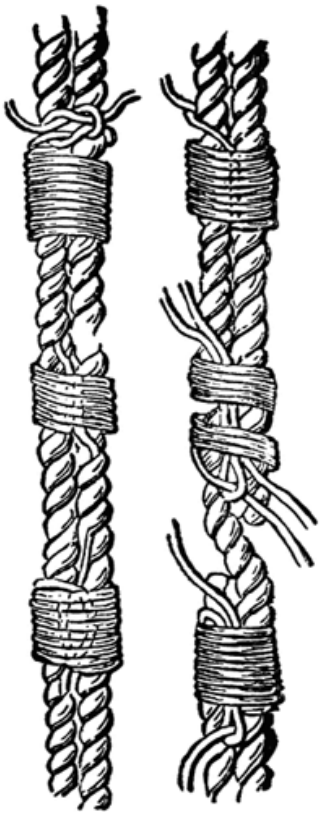
So what about the babies? Is it possible that we are again in an area where what we mean by 'conscious' and the way we carve things conceptually is half the problem? It does look a bit like it.

After all (and my apologies to any readers who may have been grinding their teeth in frustration) babies are obviously conscious, aren't they? The difference between a sleeping baby and one that is awake (which for some common-sense values, equals 'conscious') is far too salient for any parent to overlook. On the other hand, do babies soliloquise internally? Equally obviously, no, because they don't have the words to do it with.

But Kouider et al do make it fairly clear that they are specifically concerned with perception, and they make only sensible claims, noting that their results might be relevant, for example, to questions of infant anaesthesia (although it may be difficult to keep the can of phenomenal worms fully sealed on that issue). It is interesting to note the gradual development in speed and intensity which they have uncovered, but by and large I think common sense has been vindicated.

CEMI and Meaning

6 May 2013



Johnjoe McFadden has followed up the paper on his conscious electromagnetic information (CEMI) field which we discussed recently with another in the JCS - it's also featured on MLU, where you can access a copy.

This time he boldly sets out to tackle the intractable enigma of meaning. Well, actually, he says his aims are more modest; he believes there is a separate binding problem which affects meaning and he wants to show how the CEMI field offers the best way of resolving it. I think the problem of meaning is one of those issues it's difficult to sidle up to; once you've gone into the dragon's lair you tend to have to fight the beast even if all you set out to do was trim its claws; and I think McFadden is perhaps drawn into offering a bit more than he promises; nothing wrong with that, of course.

Why then, does McFadden suppose there is a binding problem for meaning? The original binding problem is to do with perception. All sorts of impulses come into our heads through different senses and get processed in different ways in different places and different speeds. Yet somehow out of these chaotic inputs the mind binds together a beautifully coherent sense of what is going on, everything matching and running smoothly with no lags or failures of lip-synch. This smoothly co-ordinated experience is robust, too; it's not easy to trip it up in the way optical illusions so readily derail up our visual processes. How is this feat pulled off? There are a range of answers on offer, including global workspaces and suggestions that the whole thing is a misconceived pseudo-problem; but I've never previously come across the suggestion that meaning suffers a similar issue.

McFadden says he wants to talk about the phenomenology of meaning. After sitting quietly and thinking about it for some time, I'm not at all sure, on the basis of introspection, that meaning has any phenomenology of its own, though no doubt when we mean things there is usually some accompanying phenomenology going on. Is there something it is like to mean something? What these perplexing words seem to portend is that McFadden, in making his case for the binding problem of meaning, is actually going to stick quite closely with perception. There is clearly a risk that he will end up talking about perception; and perception and meaning are not at all the same. For one thing the 'direction of fit' is surely different; to put it crudely, perception is primarily about the world impinging on me, whereas meaning is about me pointing at the world.

McFadden gives five points about meaning. The first is unity; when we mean a chair, we mean the whole thing, not its parts. That's true, but why is it problematic? McFadden talks about how the brain deals with impossible triangles and sees words rather than collections of letters, but that's all about perception; I'm left not seeing the problem so far as meaning goes. The second point is context-dependence. McFadden quite rightly points out that meaning is highly context sensitive and that the same sequence of letters can mean different things on different occasions. That is indeed an interesting property of meaning; but he goes on to talk about how

meanings are perceived, and how, for example, the meaning of "ball" influences the way we perceive the characters 3ALL. Again we've slid into talking about perception.

With the third point, I think we fare a bit better; this is compression, the way complex meanings can be grasped in a flash. If we think of a symphony, we think, in a sense, of thousands of notes that occur over a lengthy period, but it takes us no time at all. This is true, and it does point to some issue around parts and wholes, but I don't think it quite establishes McFadden's point. For there to be a binding problem, we'd need to be in a position where we had to start with meaning all the notes separately and then triumphantly bind them together in order to mean the symphony as a whole - or something of that kind, at any rate. It doesn't work like that; I can easily mean Mahler's eighth symphony (see, I just did it), of whose notes I know nothing, or his twelfth, which doesn't even exist.

Fourth is emergence: the whole is more than the sum of its parts. The properties of a triangle are not just the properties of the lines that make it up. Again, it's true, but the influence of perception is creeping in; when we see a triangle we know our brain identifies the lines, but we don't know that in the case of meaning a triangle we need at any stage to mean the separate lines - and in fact that doesn't seem highly plausible. The fifth and last point is interdependence: changing part of an object may change the percept of the whole, or I suppose we should be saying, the meaning. It's quite true that changing a few letters in a text can drastically change its meaning, for example. But again I don't see how that involves us in a binding problem. I think McFadden is typically thinking of a situation where we ask ourselves 'what's the meaning of this diagram?' - but that kind of example invites us to think about perception more than meaning.

In short, I'm not convinced that there is a separate binding problem affecting meaning, though McFadden's observations shed some interesting lights on the old original issue. He does go on to offer us a coherent view of meaning in general. He picks up a distinction between intrinsic and extrinsic information. Extrinsic information is encoded or symbolised according to arbitrary conventions - it sort of corresponds with derived intentionality - so a word, for example, is extrinsic information about the thing it names. Intrinsic information is the real root of the matter and it embodies some features of the thing represented. McFadden gives the following definition.¶

> *Intrinsic information exists whenever aspects of the physical relationships that exist between the parts of an object are preserved – either in the original object or its representation*.

So the word "car" is extrinsic and tells you nothing unless you can read English. A model of a car, or a drawing, has intrinsic information because it reproduces some of the relations between parts that apply in the real thing, and even aliens would be able to tell something about a car from it (or so McFadden claims). It follows that for meaning to exist in the brain there must be 'models' of this kind somewhere. (McFadden allows a little bit of wiggle room; we can express dimensions as weights, say, so long as the relationships are preserved, but in essence the whole thing is grounded in what some others might call 'iconic' representation.) Where could that be? The obvious place to look is in the neurons. but although McFadden allows that firing rates in a pattern of neurons could carry the information, he doesn't see how they can be brought together: step forward the CEMI field (though as I said previously I don't really understand why the field doesn't just smoosh everything together in an unhelpful way).

The overall framework here is sensible and it clearly fits with the rest of the theory; but there are two fatal problems for me. The first is that, as discussed above, I don't think McFadden succeeds in making the case for a separate binding problem of meaning, getting dragged back by the gravitational pull of perception. We have the original binding problem because we know perception starts with a jigsaw kit of different elements and produces a slick unity, whereas all the worries about parts seem unmotivated when it comes to meaning. If there's no new binding problem of meaning, then the appeal of CEMI as a means of solving it is obviously limited.

The second problem is that his account of meaning doesn't really cut the mustard. This is unfair, because he never said he was going to solve the whole problem of meaning, but if this part of the theory is weak it inevitably damages the rest. The problem is that representations that work because they have some of the properties of the real thing, don't really work. For one thing a glance at the definition above shows it is inherently limited to things with parts that have a physical relationship. We can't deal with abstractions at all. If I tell you I know why I'm writing this, and you ask me what I mean, I can't tell you I mean my desire for understanding, because my desire for understanding does not have parts with a physical relationship, and there cannot therefore be intrinsic information about it.

But it doesn't even work for physical objects. McFadden's version of intrinsic information would require that when I think 'car' it's represented as a specific shape and size. In discussing optical illusions he concedes at a late stage that it would be an 'idealised' car (that idealisation sounds problematic in itself); but I can mean 'car' without meaning anything ideal or particular at all. By 'car' I can in fact mean a flying vehicle with no wheels made of butter and one centimetre long (that tiny midge is going to regret settling in my butter dish as he takes his car ride into the bin of oblivion courtesy of a flick from my butter knife), something that does not in any way share parts with physical relationships which are the same as any of those applying to the big metal thing in the garage.

Attacking that flank, as I say, probably is a little unfair. I don't think the CEMI theory is going to get new oomph from the problems of meaning, but anyone who puts forward a new line of attack on any aspect of that intractable issue deserves our gratitude.

Self-Denial

19 May 2013

Consciousness, as we've noted before, is a most interdisciplinary topic, and besides the neurologists, the philosophers, the AI people, the psychologists and so on, the novelists have also, in their rigourless way, delved deep into the matter. Ever since the James boys (William and Henry) started their twin-track investigation there has been an intermittent interchange between the arts and the sciences. Academics like Dan Lloyd have written novels, novelists like our friend Scott Bakker have turned their hand to serious theory.

Recently we seem to have had a new genre of invented brain science. We could include Ian McEwan's fake paper on De Clerambault syndrome, appended to Enduring Love; recently Sebastian Faulks gave us Glockner's Isthmus; now, in his new novel A Box of Birds Charles Fernyhough gives us the Lorenzo Circuit.

The Lorenzo Circuit is a supposed structure which pulls together items from various parts of the brain and uses them to constitute memories. It's sort of assumed that the same function thereby provides consciousness and the sense of self. Since it seems unlikely that a distinct brain structure could have escaped notice this long, we must take it that the Lorenzo is a relatively subtle feature of the connectome, only identifiable through advanced scanning techniques. The Lycée, which despite its name seems to be an English university, has succeeded in mapping the circuit in detail, while Sansom, one of those large malevolent corporate entities that crop up in thrillers, has developed new electrode technology which allows safe and detailed long-term interference with neurons. It's obvious to everyone that if brought together these two discoveries would provide a potent new technology; a cure for Alzheimer's is what seems to be at the forefront of everyone's minds, though I would have thought there were far wilder and more exciting possibilities. The story revolves around the narrator, Dr Yvonne Churcher, an academic at the Lycée, and two of her undergraduate students, Gareth and James.

Unfortunately, I didn't rate the book all that highly as a novel. The plot is put together out of slightly corny thrillerish elements and seems a bit loosely managed. I didn't like the characters much either. Yvonne seems to be putty in the hands of her students, letting Gareth steal the Lycée's crucial research without seeming to hold the betrayal of her trust against him at all, and being readily seduced by the negligent James, a nonsense-talking cult member who calls her 'babe' (ack!). I've seen Gareth described as a "brilliant" character in reviews elsewhere, but sadly not much brilliance seems to be on offer. In fact to be brutal he seemed to me quite a convincing depiction of the kind of student who sits at the back of lectures chuckling to himself for no obvious reason and ultimately requires pastoral intervention. Apart from nicking other people's theories and data, his ideas seem to consist of a metaphor from Plato, which he interprets with dismal literalism.

This metaphor is the birds thing that provides the title and up to a point, the theme of the book. In the Theaetetus, Plato makes a point about how we can possess knowledge without having it

actually in our consciousness by comparing it to owning an aviary of birds without having them actually in your hand. In Plato's version there's no doubt that there's a man in the aviary who chooses the birds to catch; here I think the idea is more that he flocking and movement of the birds itself produces higher-level organisation analogous to conscious memory.

Yvonne is a pretty resolute sceptic about her own selfhood; she can't see that she is anything beyond the chance neurochemical events which sweep through her brain. This might indeed explain her apparent passivity and the way she seems to drift through even the most alarming and hare-brained adventures, though if so it's a salutary warning about the damaging potential of overdosing on materialism. Overall the book alludes to more issues than it really discusses, and gives us little side treats like a person whose existence turns out to be no more than a kind of narrative convention; perhaps it's best approached as a potential thought provoker rather than the adumbration of a single settled theory; not necessarily a bad thing for a book to be.

Yvonne's scepticism did cause me to realise that I was actually rather hazy on the subject; what is it that people who deny the self are actually denying, and are they all denying the same thing? There are actually quite a few options.¶

- I think all self-sceptics want to deny the existence of the traditional immaterial soul, and for some that may really be about all. (To digress a bit, there are actually caverns below us at this point which have not been explored for thousands of years, if ever: if we were ancient Egyptians, with their complex ontology of multiple souls, we should have a large range of sceptical permutations available; denying the ba while affirming the khaibit, say. Our simpler culture, perhaps mercifully, does not offer us such a range of refinedly esoteric entities in which to disbelieve, but those of a philosophical temperament may be inclined to cast a regretful glance towards those profoundly obscure imaginary galleries.)
- Some may want to deny any sense, or feeling, of self; like Hume they see only a bundle of sensations when they look inside themselves. I think there is arguably a quale of the self; but these people would not accept it.
- Others, by contrast, would affirm that the sense of self is vivid, just not veridical. We think there's a self, but there's nothing actually there. There's scope for an interesting discussion about what would have to be there in order to prove them wrong - or whether having the sense of self itself constitutes the self.
- Some would say that there is indeed 'something' there; it just isn't what we think it is. For example, there might indeed be a centre of experience, but an epiphenomenal one; a self who has no influence on events but is in reality just along for the ride.
- Logically I suppose we could invert that to have a self that really did make the decisions, but was deluded about having any experiences. I don't think that would be a popular option, though.
- Some would make the self a purely social construct, a matter of legal and moral rights and privileges, a conception simply grafted on to an animal which in itself, or by itself, would lack it.
- Some would deny only that the self provides a break in the natural chain of cause and effect. We are not really the origin of anything, they would say, and our impression of being a freely willing being is mistaken.
- Some radical sceptics would deny that even the body has any particular selfhood; over time every part of it changes and to assert that I am the same self as the person of twenty years ago makes no sense.

As someone who, on the whole, prefers to look for a tenable account of the reality of the self, the richness of the sceptical repertoire makes me feel rather unimaginative.

Now That's What I Call Dennett

2 June 2013

Professors are too polite. So Daniel Dennett reckons. When leading philosophers or other academics meet, they feel it would be rude to explain their theories thoroughly to each other, from the basics up. That would look as if you thought your eminent colleague hadn't grasped some of the elementary points. So instead they leap in and argue on the basis of an assumed shared understanding that isn't necessarily there. The result is that they talk past each other and spend time on profitless misunderstandings.

Dennett has a cunning trick to sort this out. He invites the professors to explain their ideas to a selected group of favoured undergraduates ('Ew; he sounds like Horace Slughorn' said my daughter); talking to undergraduates they are careful to keep it clear and simple and include an exposition of any basic concepts they use. Listening in, the other professors understand what their colleagues really mean, perhaps for the first time, and light dawns at last.

It seems a good trick to me (and for the undergraduates, yes, by 'good' I mean both clever and beneficial); in his new book Intuition Pumps and Other Tools for Thinking Dennett seems covertly to be playing another. The book offers itself as a manual or mental tool-kit offering tricks and techniques for thinking about problems, giving examples of how to use them. In the examples, Dennett runs through a wide selection of his own ideas, and the cunning old fox clearly hopes that in buying his tools, the reader will also take up his theories. (Perhaps this accessible popular presentation will even work for some of those recalcitrant profs, with whom Dennett has evidently grown rather tired of arguing.... heh, heh!)

So there's a hidden agenda, but in addition the 'intuition pumps' are not always as advertised. Many of them actually deserve a more flattering description because they address the reason, not the intuition. Dennett is clear enough that some of the techniques he presents are rather more than persuasive rhetoric, but at least one reviewer was confused enough to think that Reduction ad Absurdum was being presented as an intuition pump - which is rather a slight on a rigorous logical argument: a bit like saying Genghis Khan was among the more influential figures in Mongol society.

It seems to me, moreover, that most of the tricks on offer are not really techniques for thinking, but methods of presentation or argumentation. I find it hard to imagine someone trying to solve a problem by diligently devising thought-experiments and working through the permutations; that's a method you use when you think you know the answer and want to find ways to convince others.

What we get in practice is a pretty comprehensive collection of snippets; a sort of Dennettian Greatest Hits. Some of the big arguments in philosophy of mind are dropped as being too convoluted and fruitless to waste more time on, but we get the memorable bits of many of Dennett's best thought-experiments and rebuttals. Not all of these arguments benefit from being taken out of the context of a more systematic case, and here and there - it's inevitable I suppose - we find the remix or late cover version is less successful than the original. I thought

this was especially so in the case of the Giant Robot; to preserve yourself in a future emergency you build a wandering robot to carry you around in suspended animation for a few centuries. The robot needs to survive in an unpredictable world, so you end up having to endow it with all the characteristics of a successful animal; and you are in a sense playing the part of the Selfish Gene. Such a machine would be able to deal with meanings and intentionality just the way you do, wouldn't it? Well, in this brief version I don't really see why or, perhaps more important, how.

Dennett does a bit better with arguments against intrinsic intentionality, though I don't think his arguments succeed in establishing that there is no difference between original and derived intentionality. If Dennett is right, meaning would be built up in our brains through the interaction of gradually more meaningful layers of homunculi; OK (maybe), but that's still quite different to what happens with derived intentionality, where things get to mean something because of an agreed convention or an existing full-fledged intention.

Dennett, as he acknowledges, is not always good at following the maxims he sets out. An early chapter is given over to the rules set out by Anatol Rapoport, most notably:¶

> *You should attempt to re-express your target's position so clearly, vividly and fairly that your target says, "Thanks, I wish I'd thought of putting it that way."*

As someone on Metafilter said, when Dan Dennett does that for Christianity, I'll enjoy reading it; but there was one place in the current book where I thought Dennett fell short on understanding the opposition. He suggests that Kasparov's way of thinking about chess is probably the same as Deep Blue's in the end. What on earth could provoke one to say that they were obviously different, he protests. Wishful thinking? Fear? Well, no need to suppose so: we know that the hardware (brain versus computer) is completely different and runs a different kind of process; we know the capacities of computer and brain are different and, in spite of an argument from Dennett to the contrary, we know the heuristics are significantly different. We know that decisions in Kasparov's case involve consciousness, while Deep Blue lacks it entirely. So, maybe the processes are the same in the end, but there are some pretty good prima facie reasons to say they look very different.

One section of the book naturally talks about evolution, and there's good stuff, but it's still a twentieth century, Dawkinsian vision Dennett is trading in. Can it be that Dennett of all people is not keeping up with the science? There's no sign here of the epigenetic revolution; we're still in a world where it's all about discrete stretches of DNA. That DNA, moreover, got to be the way it is through random mutation; no news has come in of the great struggle with the viruses which we now know has left its wreckage all across the human genome, and more amazing, has contributed some vital functional stretches without which we wouldn't be what we are. It's a pity because that seems like a story that should appeal to Dennett, with his pandemonic leanings.

Still, there's a lot to like; I found myself enjoying the book more and more as it went on and the pretence of being a thinking manual dropped away a bit. Naturally some of Dennett's old attacks on qualia are here, and for me they still get the feet tapping. I liked Mr Clapgras, either a new argument or more likely one I missed first time round; he suffers a terrible event in which all his emotional and empathic responses to colour are inverted without his actual perception of colour changing at all. Have his qualia been inverted - or are they yet another layer of experience? There's really no way of telling and for Dennett the question is hardly worth asking.

When we got to Dennett's reasonable defence of compatibilism over free will, I was on my feet and cheering.

I don't think this book supersedes Consciousness Explained if you want to understand Dennett's views on consciousness. You may come away from reading it with your thinking powers enhanced, but it will be because your mental muscles have been stretched and used, not really because you've got a handy new set of tools. But if you're a Dennett fan or just like a thoughtful and provoking read, it's worth a look.

DownLoading Hauskeller

17 June 2013

Michael Hauskeller has an interesting and very readable paper in the International Journal of Machine Consciousness on uploading - the idea that we could transfer ourselves from this none-too solid flesh into a cyborg body or even just into the cloud as data. There are bits I thought were very convincing and bits I thought were totally wrong, which overall is probably a good sign.

The idea of uploading is fairly familiar by now; indeed, for better or worse it resembles ideas of transmigration, possession, and transformation which have been current in human culture for thousands of years at least. Hauskeller situates it as the logical next step in man's progressive remodelling of the environment, while also nodding to those who see it as the next step in the evolution of humankind itself. The idea that we could transfer or copy ourselves into a computer, Hauskeller points out, rests on the idea that if we recreate the right functional relationships, the phenomenological effects of consciousness will follow; that, as Minsky put it, 'Minds are what Brains do'. This remains for Hauskeller a speculation, an empirical question we are not yet in a position to test, since we have not as yet built a whole brain simulation (not sure how we would test phenomenology even after that, but perhaps only philosophers would be seriously worried about it...). In fact there are some difficulties, since it has been shown that identical syntax does not guarantee identical semantics (So two identical brains could contain identical thoughts but mean different things by them - or something strange like that. While I think the basic point is technically true with reference to derived intentionality, for example in the case of books - the same sentence written by different people can have different meanings - it's not clear to me that it's true for brains, the source of original intentionality.).

However, as Hauskeller says, uploading also requires that identity is similarly transferable, that our computer-based copy would be not just a mind, but a particular mind - our own. This is a much more demanding requirement. Hauskeller suggests the analogy of books might be brought forward; the novel Ulysses can be multiply realised in many different media, but remains the same book. Why shouldn't we be like that? Well, he thinks readers are different. Two people might both be reading Ulysses at the same moment, meaning the contents of their minds were identical; but we wouldn't say they had become the same person. Conceivably at least, the same mind could be 'read' by different selves in the same way a single book can be read by different readers.

Hauskeller's premise there is questionable - two people reading the same book don't have identical mental content (a point he has just touched on, oddly enough, since it would follow from the fact that syntax doesn't guarantee semantics, even if it didn't follow simply from the complexity of our multi-layered mental lives). I'd say the very idea of identical mental content is hard to imagine, and that by using it in thought-experiments we risk, as Dennett has warned, mistaking our own imaginative difficulties for real-world constraints. But Hauskeller's general point, that identity need not follow from content alone, is surely sound enough.

What about Ray Kurzweil's argument from gradualism? This points out that we might replace someone with cyborg parts bit by bit. We wouldn't have any doubt about the continuing identity of someone with a cyborg eye; nor someone with an electronic hippocampus. If each neuron were replaced by a functional equivalent one by one, we'd be forced to accept either that the final robot, with no biological parts at all, was indeed the same continuing person, or, that at some stage a single neuron made a stark binary difference between being the same person and not being the same person. If the final machine can be the same person, then uploading by less arduous methods is also surely possible since it's equivalent to making the final machine by another route?

Hauskeller basically bites Kurzweil's bullet. Yes, it's conceivable that at some stage there will come neurons whose replacement quite suddenly switches off the person being operated on. I have a lot of sympathy with the idea that some particular set of neurons might prove crucial to identity, but I don't think we need to accept the conceivability of sudden change in order to reject Kurzweil's argument. We can simply suppose that the subject becomes a chimera; a compound of two identically-functioning people. The new person keeps up appearances alright, but the borders of the old personality gradually shrink to destruction, though it may be very unclear when exactly that should be said to have happened.

Suppose (my example) an image of me is gradually overlaid with an image of my identical evil twin Retep, line of pixels by line. No one can even tell the process is happening, yet at some stage it ceases to be a picture of me and becomes one of Retep. The fact that we cannot tell when does not prove that I am identical with Retep, nor that both pictures are of me.

Hauskeller goes on to attack 'information idealism'. The idea of uploading often rests on the view that in the final analysis we consist of information, but¶

> *Having a mind generally means being to some extent aware of the world and oneself, and this awareness is not itself information. Rather, it is a particular way in which information is processed…*

Hauskeller, provocatively but perhaps not unjustly, accuses those who espouse information idealism of Cartesian substance dualism; they assume the mind can be separated from the body.

But no, it can't: in fact Hauskeller goes on to suggest that in fact the whole body is important to our mental life: we are not just our brains. He quotes Alva Noë and goes further, saying:¶

> *That we can manipulate the mind by manipulating the brain, and that damages to our brains tend to inhibit the normal functioning of our minds, does not show that the mind is a product of what the brain does.*

The brain might instead, he says, be like a window; if the window is obscured, we can't see beyond it, but that does not mean the window causes what lies beyond it.

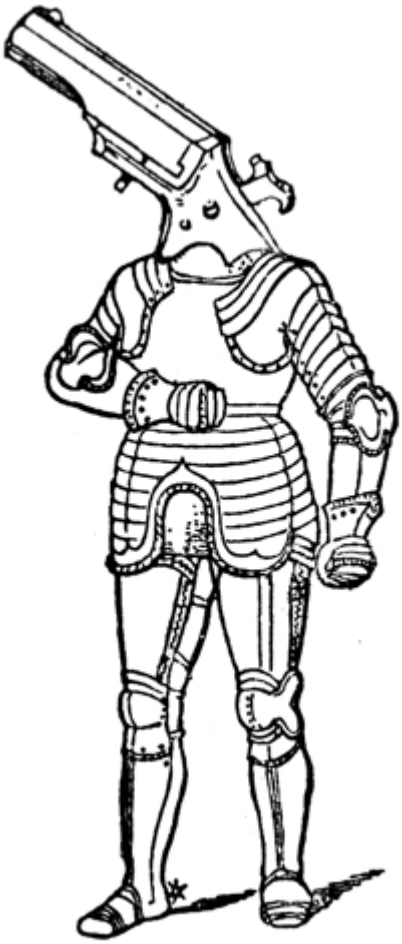
Who's sounding dualist now? I don't think that works. Suppose I am knocked unconscious by the brute physical intervention of a cosh; if the brain were merely transmitting my mind, my mental processes would continue offstage and then when normal service was resumed I should be aware that thoughts and phenomenology had been proceeding while my mere brain was disabled. But it's not like that; knocking out the brain stops mental processes in a way that blocking a window does not stop the events taking place outside.

Although I take issue with some of his reasoning, I think Hauskeller's objections have some force, and the limited conclusion he draws - that the possibility of uploading a mind, let alone an identity, is far from established, is true as far as it goes.

How much do we care about identity as opposed to continuity of consciousness? Suppose we had to chose between on the one hand retaining our bare identity while losing all our characteristics, our memories, our opinions and emotions, our intelligence, abilities and tastes, and getting instead some random stranger's equivalents; or on the other losing our identity but leaving behind a new person whose behaviour, memories, and patterns of thought were exactly like ours? I suspect some people might choose the latter.

What's Wrong With Killer Robots?

30 June 2013



There is a gathering campaign against the production and use of 'killer robots', weapons endowed with a degree of artificial intelligence and autonomy. The increasing use of drones by the USA in particular has produced a sense that this is no longer science fiction and that the issues need to be addressed before they go by default. The United Nations recently received a report proposing national moratoria, among other steps.

However, I don't feel I've yet come across a really clear and comprehensive statement of why killer robots are a problem; and it's not at all a simple matter. So I thought I'd try to set out a sketch of a full overview, and that's what this piece aims to offer. In the process I've identified some potential safguarding principles which, depending on your view of various matters, might be helpful or appropriate; these are collected together at the end.

I would be grateful for input on this - what have I missed? What have I got wrong?

In essence I think there are four broad reasons why hypothetically we might think it right to be wary of killer robots: first, because they work well; second because in other ways they don't work well, third because they open up new scope for crime, and fourth because they might be inherently unethical.

A. Because they work well.

A1. *They do bad things*. The first reason to dislike killer robots is simply that they are a new and effective kind of weapon. Weapons kill and destroy; killing and destruction is bad. The main counter-argument at an equally simple level is that in the hands of good people they will be used only for good purposes, and in particular to counter and frustrate the wicked use of other weapons. There is room for many views about who the good people may be, ranging from the view that anyone who wants such weapons disqualifies themselves from goodness automatically, to the view that no-one is morally bad and we're all equally entitled to whatever weapons we want; but we can at any rate pick up our first general principle.¶

> *P1. Killer Robots should not be produced or used in a way that allows them to fall into the hands of people who will use them unethically.*

¶A2. *They make war worse*. Different weapons have affected the character of warfare in different ways. The machine gun, perhaps, transformed the mobile engagements of the nineteenth century into the fixed slaughter of the First World War. The atomic bomb gave us the capacity to destroy entire cities at a time; potentially to kill everyone on Earth, and arguably

made total war unthinkable for rational advanced powers. Could it be that killer robots transform war in a way which is bad? There are several possible claims.

A2 (i) *They make warfare easier and more acceptable for the belligerent who has them*. No soldier on your own side is put at risk when a robot is used, so missions which are unacceptable because of the risk to your own soldier's lives become viable. In addition robots may be able to reach places or conduct attacks which are beyond the physical capacity of a human soldier, even one with special equipment. Perhaps, too, the failure of a drone does not involve a loss of face and prestige in the same way as the defeat of a human soldier. If we accept the principle that some uses of weapons are good, then for those uses this kind of objection is inverted; the risk elimination and extra capability are simply further benefits for the good user. To restrict the use of killer robots for good purposes on these grounds might then be morally wrong. Even if we think the objection is sound it does not necessarily constitute grounds for giving up killer robots altogether; instead we can merely adopt the following restriction.¶

> *P2. Killer Robots should not be used for any mission which you would not be prepared to assign to a human soldier if a human soldier were capable of executing it*.

¶A2 (ii) *They tip the balance of advantage in war in favour of established powers and governments*. Because advanced robots are technologically sophisticated and expensive, they are most easily or exclusively accessible to existing authorities and large organisations. This may make insurgency and rebellion more difficult, and that may have undemocratic consequences and strengthen the hold of tyrants. Tyrants normally have sufficient hardware to defeat rebellions in a direct confrontation anyway – it's not easy to become a tyrant otherwise. Their more serious problems come from such factors as not being able or willing to kill in the massive numbers required to defeat a large-scale popular rebellion (because that would disrupt society and hence damage their own power), disloyalty among subordinates who have control of the hardware, or inability to deal with both internal and external enemies at the same time. Killer robots make no difference to the first of these factors; they would only affect the second if they made it possible for the tyrant to dispense with human subordinates, controlling the repression-bots direct from a central control panel, or being able to safely let them get on with things at their own discretion - still a very remote contingency; and on the third they make only the same kind of difference as additional conventional arms, providing no particular reason to single out robots for restriction. However, robots might still be useful to a tyrant in less direct ways, notably by carrying out targeted surveillance and by taking out leading rebels individually in a way which could not be accomplished by human agents.

A2 (iii) *They tip the balance of advantage in war against established powers and governments* Counter to the last point, or for a different set of robots, they may help anti-government forces. In a sense we already have autonomous weapons in the shape of land-mines and IEDs, which wait, detect an enemy (inaccurately) and fire. Hobbyists are already constructing their own drones, and it's not at all hard to imagine that with some cheap or even recycled hardware it would be possible to make bombs that discriminate a little better; automatic guns that recognise the sound of enemy vehicles or the appearance of enemy soldiers, and aim and fire selectively; and crawling weapons that infiltrate restricted buildings intelligently before exploding. In A2 (ii) we thought about governments that were tyrannical, but clearly as well as justifiable rebels there are insurgents and plain terrorists whose causes and methods are wrong and wicked.

A2 (iv) *They bring war further into the civilian sphere* As a consequence of some of the presumed properties of robots – their ability to navigate complex journeys unobtrusively and detect specific targets accurately – it may be that they are used in circumstances where any other approach would bring unacceptable risks of civilian casualties. However, they may still cause collateral deaths and injuries and in general diminish the 'safe' sphere of civilian life which would in general be regarded as something to be protected, something whose erosion could have far-reaching effects on the morale and cohesion of society. Principle P2 would guard against this problem.

A3 *They circumvent ethical restrictions* The objection here is not that robots are unethical, which is discussed below under D, but that their lack of ethics makes them more effective because it enables them to do things which couldn't be done otherwise, extending the scope and consequences of war, war being inherently bad as we remember.

A3 (i) *Robots will do unethical things which could not otherwise be done.* It doesn't seem that robots can do anything that wouldn't otherwise be done for ethical reasons: either they approximate to tools, in which case they are under the control of a human agent who would presumably have done the same things without a robot had they been physically possible; or if the robots are sufficiently advanced to have real moral agency of their own, they correspond to a human agent and, subject to the discussion below, there is no reason to think they would be any better or worse than human beings.

A3 (ii) *Using robots gives the belligerent an enabling sense of having 'clean hands'*. Being able to use robots may make it easier for belligerents who are evil but squeamish to separate themselves from their actions. This might be simply because of distance: you don't have to watch the consequences of a drone attack up close; or it might be because of a genuine sense that the robot bears some of the responsibility. The greater the autonomy of the robot, the greater this latter sense is likely to be. If there are several stages in the process this sense of moral detachment might be increased: suppose we set out overall mission objectives to a strategic management computer, which then selects the resources and instructs the drones on particular tasks, leaving them with options on the final specifics and an ability to abort if required. In such circumstances it might be relatively easy to disregard the blood eventually sprayed across the landscape as a result of our instructions.

A3 (iii) *Robots can easily be used for covert missions, freeing the belligerent from the fear of punishment for unethical behaviour.* Robots facilitate secrecy both because they may be less detectable to the enemy and because they may require fewer humans to be briefed and aware, humans being naturally more inclined to leak information than a robot, especially a one-use robot.

B Because they don't work well

B1 *They are inherently likely to go wrong*. In some obvious ways robots are more reliable than human agents; their capacities can be known much more exactly and they are very predictable. However, it can be claimed that they have weaknesses, and the most important is probably inability to deal with open-ended real-life situations. AI has shown it can do well in restricted domains, but to date it has not performed well where clear parameters cannot be established in advance. This may change, but for the moment while it's quite conceivable we might send a robot after a man who fitted a certain description, it would not be a good idea to send a robot to take out anyone 'behaving like an insurgent'. This need not be an insuperable problem because

in many cases battlefield situations or the conditions of a particular mission may be predictable enough to render the use of a robot sufficiently safe; the risk may be no greater or perhaps even less, than the risk inevitably involved in using unintelligent weapons. We can guard against this systematically by adopting another principle.¶

> *P3. Killer Robots should not be used for any mission in unpredictable circumstances or where the application of background understanding may be required.*

B2 *The consequences of their going wrong are especially serious*. This is the Sorcerer's Apprentice scenario: a robot is sent out on a limited mission but for some reason does not terminate the job, and continues to threaten and kill victims indefinitely; or it destroys far more than was intended. The answer here seems to be; don't design a robot with excess killing capacity. And impose suitable limits and safeguards.¶

> *P4. Killer Robots should not be equipped with capacities that go beyond the immediate mission; they should be subject to built-in time limits and capable of being shut down remotely.*

B3 *They tempt belligerents into over-ambitious missions*. It seems plausible enough that sometimes the capacities of killer robots might be over-estimated, but that's a risk that applies to all weapons. I don't see anything about robots in themselves that makes the error especially seductive.

B4 *They lack understanding of human beings*. So do bullets, of course; the danger here only arises if we are asking the robot to make very highly sophisticated judgements about human behaviour. Robots can be equipped with game-theoretic rules that allow them to perform well in defined tactical situations; otherwise the risk is covered by principle P3. That, at least applies to current AI. It is conceivable that in future we may develop robots that actually possess 'theory of mind' in the sense required to understand human beings.

B5 *They lack emotion*. Lack of emotion can be a positive feature if the emotions are anger, spite and sadistic excitement. In general a good military agent follows rules and subject to P3, robots can do that very satisfactorily. There remains the possibility that robots will kill when a human soldier would refrain from feelings of empathy and mercy. As a countervailing factor the human soldier need not be doing the best thing, of course, and short-term immediate mercy might in some circumstances lead to more deaths overall. I believe that there are in practice few cases of soldiers failing to carry out a bad mission through feelings of sympathy and mercy, though I have no respectable evidence either way. I mentioned above the possibility that robots may eventually be endowed with theory of mind. Anything we conclude about this must be speculative to some degree, but there is a possibility that the acquisition of theory of mind requires empathy, and if so we could expect that robots capable of understanding human emotion would necessarily share it. It need not be a permanent fact that robots lack emotion. While current AI simulation of emotion is not generally impressive or useful, that may not remain the case

C Because they create new scope for crime

C1 *They facilitate 'rational' crime*. We tend to think first of military considerations, but it is surely the case that killer robots, whether specially designed or re-purposed from military uses could be turned to crime. The scope for use in murder, robbery and extortion is obvious.

C2 *They facilitate irrational crime*. A particularly unpleasant prospect is the possibility of autonomous weapons being used for irrational crime - mass murder, hate crime and so on. When computer viruses became possible, people created them even though thre was no benefit involved; it cannot be expected that people will refrain from making murder-bots merely because it makes no sense.

D Because they are inherently unethical

D1 *They lack ethical restraint*. If ethics involves obeying a set of rules, then subject to their being able to understand what is going on, robots should in that limited sense be ethical. If soldiers are required to apply utilitarian ethics, then robots at current levels of sophistication will be capable of applying Benthamite calculus, but have great difficulty identifying the values to be applied. Kantian ethics requires one to have a will and desires of one's own, so pending the arrival of robots with human-level cognition, they are presumably non-starters at it, as they would presumably be for non-naturalistic or other ethical systems. But we're not requiring robots to behave well, only to avoid behaving badly - I don't think anything beyond obedience to a set of rules is generally required because if the rules are drawn conservatively it should be possible to avoid the grey areas.

D2 *They disperse or dispel moral responsibility*. Under A3(ii) I considered the possibility that using robots might give a belligerent a false sense of moral immunity; but what if robots really do confer moral immunity? Isaac Asimov gives an example of a robot subject to a law that it may not harm human beings, but without a duty to protect them. It could, he suggests, drop a large weight towards a human being: because it knows it has ample time to stop the weight this in itself does not amount to harming the human. But once the weight is on its way, the robot has no duty to rescue the human being and can allow the weight to fall. I think it is a law of ethics that responsibility ends up somewhere except in cases of genuine accident; either with the designer, the programmer, the user, or, if sufficiently sophisticated, the robot itself. Asimov's robot equivocates; dropping the weight is only not murder in the light of a definite intention to stop the weight, and changing its mind subsequently amounts to murder.

D3 *They may become self-interested*. The classic science fiction scenario is that robots develop interests of their own and Kill All Humans in order to protect them. This could only be conceivable in the still very remote contingency that robots were endowed with full human-level capacities for volition and responsibility. If that were the case, then robots would arguably have a right to their own interests in just the same way as any other sentient being; there's no reason to think they would be any more homicidial in pursung them than human being themselves.

D4 *They resemble slaves*. The other side of the coin then, is that if we had robots with full human mental capacity, they would amount to our slaves, and we know that slavery is wrong. That isn't necessarily so, we could have killer robots that were citizens in good standing; however all of that is still a very remote prospect. Of more immediate concern is the idea that having mechanical slaves would lead us to treat human beings more like machines. A commander who gets used to treating his drones as expendable might eventually begin to be more careless with human lives. Against his there is a contrary argument that being relieved from the need to accept that war implied death for some on one's own soldiers would mean it in fact became even more shocking and less acceptable in future.

D5 *It is inherently wrong for a human to be killed by a machine*. Could it be that there is something inherently undignified or improper about being killed by a mechanical decision, as

opposed to a simple explosion? People certainly speak of it's being repugnant that a machine should have the power of life or death over a person. I fear there's some equivocation going on here: either the robot is a simple machine, in which case no moral decision is being made and no power of life or death is being exercised; or the robot is a thinking being like us, in which case it seems like mere speciesism to express repugnance we wouldn't feel for a human being in the same circumstances. I think nevertheless that even if we confine ourselves to non-thinking robots there may perhaps be a residual sense of disgust attached to killing by robot, but it does not seem to have a rational basis, and again the contrary argument can be made; that death by robot is 'clean'. To be killed by a robot seems in one way to carry some sense of humiliation, but in another sense to be killed by a robot is to be killed by no-one, to suffer a mere accident.

What conclusion can we reach? There are at any rate some safeguards that could be put in place on a precautionary basis. Beyond that I think one's verdict rests on whether the net benefits, barely touched on here, exceed the net risks; whether the genie is already out of the bottle, and if so whether it is allowable or even a duty to try to ensure that the good people maintain parity of fire-power with the bad.¶

> *P1. Killer Robots should not be produced or used in a way that allows them to fall into the hands of people who will use them unethically.*
>
> *P2. Killer Robots should not be used for any mission which you would not be prepared to assign to a human soldier if a human soldier were capable of executing it.*
>
> *P3. Killer Robots should not be used for any mission in unpredictable circumstances or where the application of background understanding may be required.*
>
> *P4. Killer Robots should not be equipped with capacities that go beyond the immediate mission; they should be subject to built-in time limits and capable of being shut down remotely.*

A General Taxonomy of Lust

14 July 2013

That's not really what this piece is about (sorry). It's an idea I had years ago for a short story or a novella. 'Lust' here would have been interpreted broadly as any state which impels a human being towards sex. I had in mind a number of axes defining a general 'lust space'. One of the axes, if I remember rightly, had specific attraction to one person at one end and generalised indiscriminate enthusiasm at the other; another went from sadistic to masochistic, and so on. I think I had eighty-one basic forms of lust, and the idea was to write short episodes exemplifying each one: in fact, to interweave a coherent narrative with all of them in.

My creative gifts were not up to that challenge, but I mention it here because one of the axes went from the purely intellectual to the purely physical. At the intellectual extreme you might have an elderly homosexual aristocrat who, on inheriting a title, realises it is his duty to attempt to procure an heir. At the purely physical end you might have an adolescent boy on a train who notices he has an erection which is unrelated to anything that has passed through his mind.

That axis would have made a lot of sense (perhaps) to Luca Barlassina and Albert Newen, whose paper in Philosophy and Phenomenological Research sets out an impure somatic theory of the emotions. In short, they claim that emotions are constituted by the integration of bodily perceptions with representations of external objects and states of affairs.

Somatic theories say that emotions are really just bodily states. We don't get red in the face because we're angry, we get angry because we've become red in the face. As no less an authority than William James had it:¶

The more rational statement is that we feel sorry because we cry, angry because we strike, afraid because we tremble, and not that we cry, strike, or tremble, because we are sorry, angry, or fearful, as the case may be. Without the bodily states following on the perception, the latter would be purely cognitive in form, pale, colorless, destitute of emotional warmth.

¶This view did not appeal to everyone, but the elegantly parsimonious reduction it offers has retained its appeal, and Jesse Prinz has put forward a sophisticated 21st century version. It is Prinz's theory that Barlassina and Newen address; they think it needs adulterating, but they clearly want to build on Prinz's foundations, not reject them.

So what does Prinz say? His view of emotions fits into the framework of his general view about perception: for him, a state is a perceptual state if it is a state of a dedicated input system - eg the visual system. An emotion is simply a state of the system that monitors our own bodies; in other words emotions are just perceptions of our own bodily states. Even for Prinz, that's a little too pure: emotions, after all, are typically about something. They have intentional content. We don't just feel angry, we feel angry about something or other. Prinz regards emotions as having dual content: they register bodily states but also represent core relational themes (as against say, fatigue, which both registers and represents a bodily state). On top of that, they may involve

propositional attitudes, thoughts about some evocative future event, for example, but the propositional attitudes only evoke the emotions, they don't play any role in constituting them. Further still, certain higher emotions are recalibrati0ns of lower ones: the simple emotion of sadness is recalibrated so it can be controlled by a particular set of stimuli and become guilt.

So far so good. Barlassina and Newen have four objections. First, if Prinz is right, then the neural correlates of emotion and the perception of the relevant bodily states must just be the same. Taking the example of disgust, B&N argue that the evidence suggests otherwise: interoception, the perception of bodily changes, may indeed cause disgust, but does not equate to it neurologically.

Second, they see problems with Prinz's method of bringing in intentional content. For Prinz emotions differ from mere bodily feeling because they represent core relational themes. But, say B&N, what about ear pressure? It tells us about unhealthy levels of barometric pressure and oxygen, and so relates to survival, surely a core relational theme: and it's certainly a perception of a bodily state - but ear pressure is not an emotion.

Third, Prinz's account only allows emotions to be about general situations; but in fact they are about particular things. When we're afraid of a dog, we're afraid of that dog, we're not just experiencing a general fear in the presence of a specific dog.

Fourth, Prinz doesn't fully accommodate the real phenomenology of emotions. For him, fear of a lion is fear accompanied by some beleifs about a lion: but B&N maintain that the directedness of the emotion is built in, part of the inherent phenomenology.

Barlassina and Newen like Prinz's somatic leanings, but they conclude that he simply doesn't account sufficiently for the representative characteristics of emotions: consequently they propose an 'impure' theory by which emotions are cognitive states constituted when interoceptive states are integrated with with perceptions of external objects or states of affairs.

This pollution or elaboration of the pure theory seems pretty sensible and B&N give a clear and convincing exposition. At the end of the day it leaves me cold not because they haven't done a good job but because I suspect that somatic theories are always going to be inadequate: for two reasons.

First, they just don't capture the phenomenology. There's no doubt at all that emotions are often or typically characterised or coloured by perception of distinctive bodily states, but is that what they are in essence? It doesn't seem so. It seems possible to imagine that I might be angry or sad without a body at all: not, of course, in the same good old human way, but angry or sad nevertheless. There seems to be something almost qualic about emotions, something over and above any of the physical aspects, characteristic though they may be.

Second, surely emotions are often essentially about dispositions to behave in a certain way? An account of anger which never mentions that anger makes me more likely to hit people just doesn't seem to cut the mustard. Even William James spoke of striking people. In fact, I think one could plausibly argue that the physical changes associated with an emotion can often be related to the underlying propensity to behave in a certain way. We begin to breathe deeply and our heart pounds because we are getting ready for violent exertion, just as parallel cognitive changes get us ready to take offence and start a fight. Not all emotions are as neat as this: we've talked in the past about the difficulty of explaining what grief is for. Still, these considerations

seem enough to show that a somatic account, even an impure one, can't quite cover the ground.

Still, just as Barlassina and Newen built on Prinz, it may well be that they have provided some good foundation work for an even more impure theory.

Turing Test Tactics

28 July 2013

2012 was Alan Turing Year, marking the hundredth centenary of his birth. The British Government, a little late perhaps, announced recently that it would support a Bill giving Turing a posthumous pardon; Gordon Brown, then the Prime Minister, had already issued an official apology in 2009. As you probably know, Turing, who was gay, was threatened with blackmail by one of his lovers (homosexuality being still illegal at the time) and reported the matter to the authorities; he was then tried and convicted and offered a choice of going to jail or taking hormones, effectively a form of oestrogen. He chose the latter, but subsequently died of cyanide poisoning in what is generally believed to have been suicide, leaving by his bed a partly-eaten apple, thought by many to be a poignant allusion to the story of Snow White. In fact it is not clear that the apple had any significance or that his death was actually suicide

The pardon was widely but not universally welcomed: some thought it an empty gesture; some asked why Turing alone should be pardoned; and some even saw it as an insult, confirming by implication that Turing's homosexuality was indeed an offence that needed to be forgiven.

Turing is generally celebrated for wartime work at Bletchley Park, the code-breaking centre, and for his work on the Halting Problem: on the latter he was pipped at the post by Alonzo Church, but his solution included the elegant formalisation of the idea of digital computing embodied in the Turing Machine, recognised as the foundation stone of modern computing. In a famous paper from 1950 he also effectively launched the field of Artificial Intelligence, and it is here that we find what we now call the Turing Test, a much-debated proposal that the ability of machines to think might be tested by having a short conversation with them.

Turing's optimism about artificial intelligence has not been justified by developments since: he thought the Test would be passed by the end of the twentieth century. For many years the Loebner Prize contest has invited contestants to provide computerised interlocutors to be put through a real Turing Test by a panel of human judges, who attempt to tell which of their conversational partners, communicating remotely by text on a screen, is human and which machine. None of the 'chat-bots' has succeeded in passing itself off as human so far - but then so far as I can tell none of the candidates ever pretended to be a genuinely thinking machine - they're simply designed to scrape through the test by means of various cunning tricks - so according to Turing, none of them should have succeeded.

One lesson which has emerged from the years of trials - often inadvertently hilarious - is that success depends strongly on the judges. If the judge allows the chat-bot to take the lead and steer the conversation, a good impression is liely to be possible; but judges who try to make things difficult for the computer never fail. So how do you go about tripping up a chat-bot?

Well, we could try testing its general knowledge. Human beings have a vast repository of facts, which even the largest computer finds it difficult to match. One problem with this approach is that human beings cannot be relied on to know anything in particular - not knowing the year of the battle of Hastings, for example, does not prove that you're not human. The second problem is that computers have been getting much better at this. Some clever chat-bots these days are

permanently accessible online; they save the inputs made by casual visitors and later discreetly feed them back to another subject, noting the response for future use. Over time they accumulate a large database of what humans say in these circumstances and what other humans say in response. The really clever part of this strategy is that not only does it provide good responses, it means your database is automatically weighted towards the most likely topics and queries. It turns out that human beings are fairly predictable, and so the chat-bot can come back with responses that are sometimes eerily good, embodying human-style jokes, finishing quotations, apparently picking up web-culture references, and so on.

If we're subtle we might try to turn this tactic of saving real human input against the chat-bot, looking for responses that seem more appropriate for someone speaking to a chat-bot than someone engaging in normal conversation, or perhaps referring to earlier phases of the conversation that never happened. But this is a tricky strategy to rely on, generally requiring some luck.

Perhaps rather than trying established facts, it might be better to ask the chat-bot questions which have never been asked before in the entire history of the world, but which any human can easily answer. When was the last time a mole fought an octopus? How many emeralds were in the crown worn by Shakespeare during his visit to Tashkent?

It might be possible to make things a little more difficult for the chat-bot by asking questions that require an answer in a specific format; but it's hard to do that effectively in a Turing Test because normal usage is generally extremely flexible about what it will accept as an answer; and failing to match the prescribed format might be more human rather than less. Moreover, rephrasing is another field where the computers have come on a lot: we only have to think of the Watson system's performance at the quiz game Jeopardy, which besides rapid retrieval of facts required just this kind of reformulation.

So it might be better to move away from general stuff and ask the chat-bot about specifics that any human would know but which are unlikely to be in a database - the weather outside, which hotel it is supposedly staying at. Perhaps we should ask it about its mother, as they did in similar circumstances in Blade Runner, though probably not for her maiden name.

On a different tack, we might try to exploit the weakness of many chat-bots when it comes to holding a context: instead of falling into the standard rhythm of one input, one response, we can allude to something we mentioned three inputs ago. Although they have got a little better, most chat-bots still seem to have great difficulty maintaining a topic across several inputs or ensuring consistency of response. Being cruel, we might deliberately introduce oddities that the bot needs to remember: we tell it our cat is called Fish and then a little later ask whether it thinks the Fish we mentioned likes to swim.

Wherever possible we should fall back on Gricean implicature and provide good enough clues without spelling things out. Perhaps we could observe to the chat-bot that poor grammar is very human - which to a human more or less invites an ungrammatical response, although of course we can never rely on a real human's getting the point. The same thing is true, alas, in the case of some of the simplest and deadliest strategies, which involve changing the rules of discourse. We tell the chat-bot that all our inputs from now on lliw eb delleps tuo sdrawkcab and ask it to reply in the same way, or we js mss t ll th vwls.

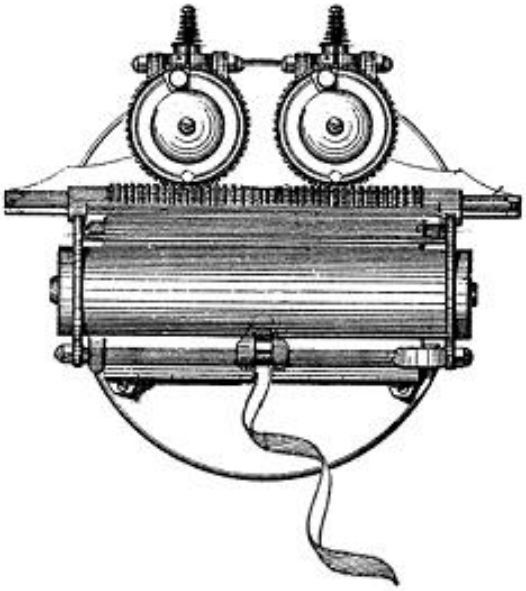
Devising these strategies makes us think in potentially useful ways about the special qualities of human thought. If we bring all our insights together, can we devise an Ultra-Turing Test? That

would be a single question which no computer ever answers correctly and all reasonably alert and intelligent humans get right. We'd have to make some small allowance for chance, as there is obviously no answer that couldn't be generated at random in some tiny number of cases. We'd also have to allow for the fact that as soon as any question was known, artful chat-bot programmers would seek to build in an answer; the question would have to be such that they couldn't do that successfully.

Perhaps the question would allude to some feature of the local environment which would be obvious but not foreseeable (perhaps just the time?) but pick it out in a non-specific allusive way which relied on the ability to generate implications quickly from a vast store of background knowledge. It doesn't sound impossible…

Computation Unlimited

10 August 2013



Massimo Pigliucci issued a spirited counterblast to computationalism recently, which I picked up on MLU. He says that people too often read the Turing-Church hypothesis as if it said that a Universal Turing Machine could do anything that any machine could do. They then take that as a basis on which to help themselves to computationalism. He quotes Jack Copeland as saying that a myth has arisen on the matter, and citing examples where he feels that Dennett and the Copelands have mis-stated the position. Actually, says Pigliucci, Turing merely tells us that a Universal Turing Machine can do anything a specific Turing machine can do, and that does not tell us what real-world machines can or can't do.

It's possible some nits are being picked here. Copeland's reported view seems a trifle too puritanical in its refusal to look at wider implications; I think Turing himself would have been surprised to hear that his work told us nothing about the potential capacities of real world digital computers. But of course Pigliucci is quite right that it doesn't establish that the brain is computational. Indeed, Turing's main point was arguably about the limits of computation, showing that there are problems that cannot be handled computationally. It's sort of part of our bedrock understanding of computation that there are many non-computable problems; apart from the original halting problem the tiling problem may be the most familiar. Tiling problems are associated with the ingenious work of Roger Penrose, and he, of course, published many years ago now what he claims is a proof that when mathematicians are thinking original mathematical thoughts they are not computing.

So really Pigliucci's moderate conclusion that computationalism remains an open issue ought to be uncontroversial? Surely no-one really supposes that the debate is over? Strangely enough there does seem to have been a bit of a revival in hard-line computationalism. Pigliucci goes on to look at pancomputationalism, the view that every natural process is instantiating a computation (or even all possible computations. This is rather like the view John Searle once proposed, that a window can be seen as a computer because it has two states, open and closed, which are enough to express a stream of binary digits. I don't propose to go into that in any detail, except to say I think I broadly agree with Pigliucci that it requires an excessively liberal use of interpretation. In particular, I think in order to interpret everything as a computation, we generally have to allow ourselves to interpret the same physical state of the object as different computational states at different times, and that's not really legitimate. If I can do that I can interpret myself into being a wizard, because I'm free to interpret my physical self as human at one time, a dragon at another, and a fluffy pink bunny at a third.

But without being pancomputationalists we might wonder why the limits of computation don't hit us in the face more often. The world is full of non-computable problems, but they rarely seem to give us much difficulty. Why is that? One answer might be in the amusing argument put by Ray Kurzweil in his book How to Create a mind. Kurzweil espouses a doctrine called the

“Universality of Computation” which he glosses as ‘the concept that a general-purpose computer can implement any algorithm”. I wonder whether that would attract a look of magisterial disapproval from Jack Copeland? Anyway, Kurzweil describes a non-computable problem known as the ‘busy beaver’ problem. The task here is to work out for a given value of n what the maximum number of ones written by any Turing machine with n states will be. The problem is uncomputable in general because as the computer (a Universal Turing Machine) works through the simulation of all the machines with n states, it runs into some that get stuck in a loop and don’t halt.

So, says Kurzweil, an example of the terrible weakness of computers when set against the human mind? Yet for many values of n it happens that the problem is solvable, and as a matter of fact computers have solved many such particular cases - many more than have actually been solved by unaided human thought! I think Turing would have liked that; it resembles points he made in his famous 1950 essay on Computing Machinery and Intelligence.

Standing aside from the fray a little, the thing that really strikes me is that the argument seems such a blast from the past. This kind of thing was chewed over with great energy twenty or even thirty years ago, and in some respects it doesn’t seem as important as it used to. I doubt whether consciousness is purely computational, but it may well be subserved, or be capable of being subserved, by computational processes in important ways. When we finally get an artificial consciousness, it wouldn’t surprise me if the heavy lifting is done by computational modules which either relate in a non-computational way or rely on non-computational processing, perhaps in pattern recognition, though Kurzweil would surely hate the idea that that key process might not be computed. I doubt whether the proud inventor on that happy day will be very concerned with the question of whether his machine is computational or not.

Experience and Autonomy

26 August 2013

Tom Clark has an interesting paper on Experience and Autonomy: Why Consciousness Does and Doesn't Matter, due to appear as a chapter in Exploring the Illusion of Free Will and Responsibility (if your heart sinks at the idea of discussing free will one more time, don't despair: this is not the same old stuff).

In essence Clark wants to propose a naturalised conception of free will and responsibility and he seeks to dispel three particular worries about the role of consciousness; that it might be an epiphenomenon, a passenger along for the ride with no real control; that conscious processes are not in charge, but are subject to manipulation and direction by unconscious ones; and that our conception of ourselves as folk-dualist agents, able to step outside the processes of physical causation but still able to intervene in them effectively, is threatened. He makes it clear that he is championing phenomenal consciousness, that is, the consciousness which provides real if private experiences in our minds; not the sort of cognitive rational processing that an unfeeling zombie would do equally well. I think he succeeds in being clear about this, though it's a bit of a challenge because phenomenal consciousness is typically discussed in the context of perception, while rational decision-making tends to be seen in the context of the 'easy problem' - zombies can make the same decisions as us and even give the same rationales. When we talk about phenomenal consciousness being relevant to our decisions, I take it we mean something like our being able to sincerely claim that we 'thought about' a given decision in the sense that we had actual experience of relevant thoughts passing through our minds. A zombie twin would make identical claims but the claims would, unknown to the zombie, be false, a rather disturbing idea.

I won't consider all of Clark's arguments (which I am generally in sympathy with), but there are a few nice ones which I found thought-provoking. On epiphenomenalism, Clark has a neat manoeuvre. A commonly used example of an epiphenomenon, first proposed by Huxley, is the whistle on a steam locomotive; the boiler, the pistons, and the wheels all play a part in the causal story which culminates in the engine moving down the track; the whistle is there too, but not part of that story. Now discussion has sometimes been handicapped by the existence of two different conceptions of epiphenomenalism; a rigorous one in which there really must be no causal effects at all, and a looser one in which there may be some causal effects but only ones that are irrelevant, subliminal, or otherwise ignorable. I tend towards the rigorous conception myself, and have consequently argued in the past that the whistle on a steam engine is not really a good example. Blowing the whistle lets steam out of the boiler which does have real effects. Typically they may be small, but in principle a long enough blast can stop a train altogether.

But Clark reverses that unexpectedly. He argues that in order to be considered an epiphenomenon an entity has to be the sort of thing that might have had a causal role in the process. So the whistle is a good example; but because consciousness is outside the third-person account of things altogether, it isn't even a candidate to be an epiphenomenon!

Although that inverts my own outlook, I think it's a pretty neat piece of footwork. If I wanted a come-back I think I would let Clark have his version of epiphenomenalism and define a new kind, x-epiphenomenalism, which doesn't require an entity to be the kind of thing that could have a causal role; I'd then argue that consciousness being x-epiphenomenal is just as worrying as the old problem. No doubt Clark in turn might come back and argue that all kinds of unworrying things were going to turn out to be x-epiphenomenal on that basis, and so on; however, since I don't have any great desire to defend epiphenomenalism I won't even start down that road.

On the second worry Clark gives a sensible response to the issues raised by the research of Libet and others which suggest our decisions are determined internally before they ever enter our consciousness; but I was especially struck by his arguments on the potential influence of unconscious factors which form an important part of his wider case. There is a vast weight of scientific evidence to show that often enough our choices are influenced or even determined by unconscious factors we're not aware of; Clark gives a few examples but there are many more. Perhaps consciousness is not the chief executive of our minds after all, just the PR department?

Clark nibbles the bullet a bit here, accepting that unconscious influence does happen, but arguing that when we are aware of say, ethnic bias or other factors, we can consciously fight against it and second-guess our unworthier unconscious impulses. I like the idea that it's when we battle our own primitive inclinations that we become most truly ourselves; but the issues get pretty complicated.

As a side issue, Clark's examples all suppose that more or less wicked unconscious biases are to be defeated by a more ethical conscious conception of ourself (rather reminiscent of those cartoon disputes between an angel on the character's right shoulder and a devil on the left); but it ain't necessarily so. What if my conscious mind rules out on principled but sectarian grounds a marriage to someone I sincerely love with my unconscious inclinations? I'm not clear that the sectarian is to be considered the representative of virtue (or of my essential personal agency) more than the lover.

That's not the point at all, of course: Clark is not arguing that consciousness is always right, only that it has a genuine role. However, the position is never going to be clear. Suppose I am inclined to vote against candidate N, who has a big nose. I tell myself I should vote for him because it's the schnozz that is putting me off. Oh no, I tell myself, it's his policies I don't like, not his nose at all. Ah, but you would think that, I tell myself, you're bound to be unaware of the bias, so you need to aim off a bit. How much do \\I aim off, though - am I to vote for all big-nosed candidates regardless? Surely I might also have legitimate grounds for disliking them? And does that 'aiming off' really give my consciousness a proper role or merely defer to some external set of rules?

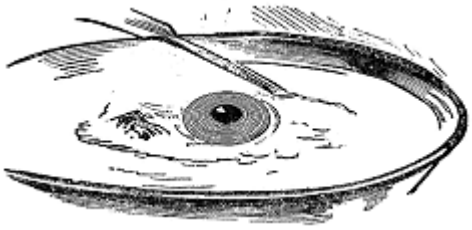
Worse yet, as I leave the polling station it suddenly occurs to me that the truth is, the nose had nothing to do with it; I really voted for N because I'm biased in favour of white middle-aged males; my unconscious fabricated the stuff about the nose to give me a plausible cover story while achieving its own ends. Or did it? Because the influences I'm fighting are unconscious, how will I ever know what they really are, and if I don't know, doesn't the claimed role of consciousness become merely a matter of faith? It could always turn out that if I really knew what was going on, I'd see my consciousness was having its strings pulled all the time. Consciousness can present a rationale which it claims was effective, but it could do that to begin with; it never knew the rationale was really a mask for unconscious machinations.

The last of the three worries tackled by Clark is not strictly a philosophical or scientific one; we might well say that if people's folk-dualist ideas are threatened, so much the worse for them. There is, however, some evidence that undiluted materialism does induce what Clark calls a "puppet" outlook in which people's sense of moral responsibility is weakened and their behaviour worsened. Clark provides rational answers but his views tend to put him in the position of conceding that something has indeed been lost. Consciousness does and doesn't matter. I don't think anything worth having can be lost by getting closer to the truth and I don't think a properly materialist outlook is necessarily morally corrosive - even in a small degree. I think what we're really lacking for the moment is a sufficiently inspiring, cogent, and understood naturalised ethics to go with our naturalised view of the mind. There's much to be done on that, but it's far from hopeless (as I expect Clark might agree).

There's much more in the paper than I have touched on here; I recommend a look at it.

Conscious Yogurt

8 September 2013



Can machines be moral agents? There's a bold attempt to clarify the position in the ¶IJMC , by Parthemore and Whitby (draft version). In essence they conclude that the factors that go to make a moral agent are the same whether the entity in question is biological or an artefact. On the face of it that seems like common sense - although the opposite conclusion would be more interesting; and there is at least one goodish reason to think that there's a special problem for artefacts.

But let's start at the beginning. Parthemore and Whitby propose three building blocks which any successful candidate for moral agency must have: the concept of self, the concept of morality, and the concept of concept.

By the concept of self, they mean not simply a basic awareness of oneself as an object in the landscape, but a self-reflective awareness along the lines of Damasio's autobiographical self. Their rationale is not set out very explicitly, but I take it they think that without such a sense of self, your acts would seem to be no different from other events in the world; just things you notice happening, and that therefore you couldn't be seen as responsible for them. It's a reasonable claim, but I think it properly requires more discussion. A sceptic might make the case that there's a difference between feeling you're not responsible and actually not being responsible; that someone could have the cheerily floaty feeling of being a mere observer while actually carrying out acts that were premeditated in some level of the mind. There's scope then for an extensive argument about whether conscious awareness of making a decision is absolutely necessary to moral ownership of the decision. We'll save that for another day, and of course Parthemore and Whitby couldn't hope to deal with every implication in a single paper. I'm happy to take their views on the self as reasonable working assumptions.

The second building block is the concept of morality. Broadly, Parthemore and Whitby want their moral agent to understand and engage with the moral realm; it's not enough, they say, to memorise by rote learning a set of simple rules. Now of course many people have seemed to believe that learning and obeying a set of rules such as the Ten Commandments was necessary or even sufficient for being a morally good person. What's going on here? I think this becomes clearer when we move on to the concept of concept, which roughly means that the agent must understand what they're doing. Parthemore and Whitby do not mean that a moral agent must have a philosophical grasp of the nature of concepts, only that they must be able to deal with them in practical situations, generalising appropriately where necessary. So I think what they're getting at is not that moral rules are invalid in themselves, merely that a moral agent has to have sufficient conceptual dexterity to apply them properly. A rule about not coveting your neighbour's goods may be fine, but you need to able to recognise neighbours and goods and instances of their being coveted, without needing say, a list of the people to be considered neighbours.

So far, so good, but we seem to be missing one item normally considered fundamental: a capacity for free action. I can be fully self-aware, understand and appreciate that stealing is wrong, and be aware that by picking up a chocolate bar without paying I am in fact stealing; but

it won't generally be considered a crime if I have a gun to my head, or have been credibly told that if I don't steal the chocolate several innocent people will be massacred. More fundamentally, I won't be held responsible if it turns out that because of internal factors I have no ability to choose otherwise: yet the story told by physics seems to suggest I never really have the ability to choose otherwise. I can't have real responsibility without real free will (or can I?).

Parthemore and Whitby don't really acknowledge this issue directly; but they do go on to add what is effectively a fourth requirement for moral agency; you have to be able to act against your own interests. It may be that this is in effect their answer; instead of a magic-seeming capacity for free will they call for a remarkable but fully natural ability to act unselfishly. They refer to this as akrasia; consciously doing the worse thing: normally I think the term refers to the inability to do what you can see is the morally right thing; here Parthemore and Whitby seem to reinterpret it as the ability to do morally good things which run counter to your own selfish interests.

There are a couple of issues with that principle. First, it's not actually the case that we only act morally when going against our own interests; it's just that those are the morally interesting cases because we can be sure in those instances that morality alone is the motivation. Worse than that, someone like Socrates would argue that moral action is always in your own best interests, because being a good person is vastly more important than being rich or successful; no rational person who understood the situation properly would ever choose to do the wrong thing. Probably though, Parthemore and Whitby are really going for something a little different. They link their idea to personal boundaries, citing Andy Clark, so I think they have in mind an ability to extend sympathy or a feeling of empathy to others. The ability they're after is not so much that of acting against your own interests as that of construing your own interests to include those of other entities.

Anyway, having set out a conception of moral agency, are artefacts capable of qualifying? Parthemore and Whitby cite an thought-experiment of Zlatev: suppose someone who lived a normal life were found after death to have had no brain but a small mechanism in their skull: would we on that account disavow the respect and friendship we might have felt for the person during life? Zlatev suggests not; and Parthemore and Whitby, agreeing, propose that it would make no difference if the skull were found to be full of yogurt; indeed, supposing someone who had been found to have yogurt instead of brains were able to continue their life, they see no reason to treat them any differently on account their acephaly (galactocephaly?). They set this against John Searle's view that it is some as-yet-unidentified property of nervous tissue that generates consciousness, and that a mind made out of beer cans is a patent absurdity. Their view, which certainly has its appeal, is that it is performance that matters; if an entity displays all the signs of moral sense, then let's treat it as a moral being.

Here again Parthemore and Whitby make a reasonable case but seem to neglect a significant point. The main case against artefacts being agents is not the the Searlian view, but a claim that in the case of artefacts responsibility devolves backwards to the person who designed them, who either did or should have foreseen how they would behave; is at any rate responsible for their behaving as they do; and therefore bears any blame. My mother is not responsible for my behaviour because she did not design me or program my brain, but the creator of Robot Peter would not have the same defence. It may be that in Parthemore and Whitby's view akrasia takes care of this too, but if so it needs explaining.

If Parthemore and Whitby think performance is what matter you might think they would be well-disposed towards a Moral Turing Test; one in which the candidate's ability to discuss ethical

issues coherently determines whether we should see it as a moral agent or not. Just such a test was proposed by Allen et al, but in fact Parthemore and Whitby are not keen on it. For one thing, as they point out, it requires linguistic ability, whereas they want moral agency to extend to at least some entities with no language competence. Perhaps it would be possible to devise pre-linguistic tests, but they foresee difficulties: rightly, I think. One other snag with a Moral Turing Test would be the difficulty of spotting cases where the test subject had a valid system of ethics which nevertheless differed from our own.

The paper goes on to describe conceptual space theory and its universal variant: an ambitious proposal to map the whole space of ideas in a way which the authors think might ground practical models of moral agency. I admire the optimism of the project, but doubt whether any such mapping is possible. Tellingly, the case of colour space, which does lend itself beautifully to a simple 3d mapping, is quoted: I think other parts of the conceptual repertoire are likely to be much more challenging. Interestingly I thought the general drift suggested that the idea was a cousin of Arnold Trehub's retinoid theory, though more conceptual and perhaps not as well rooted in neurology.

Overall it's an interesting paper: Parthemore and Whitby very reasonably say at several points that they're not out to solve all the problems of philosophy; but I think if they want their points to stick they will unavoidably need to delve more deeply in a couple of places.

Third Consciousness

22 September 2013

There have been a number of reports of a speech or presentation by Dr. Jaideep Pandit to the Annual Congress of the The Association of Anaesthetists of Great Britain and Ireland (AAGBI) in Dublin, in which he apparently proposed the existence of a 'third' state of consciousness.

Dr Pandit led the fifth annual survey (NAP5) by the AAGBI on the subject of accidental awareness, in which patients who have been anaesthetised for an operation become conscious during the procedure but are unable to do anything because some of the drugs they are normally given induce paralysis. Anaesthetists used to believe that even though patients couldn't move, it was generally possible to tell when they were becoming aware, through signs of distress like the heartrate and sweating, allowing the anaesthetist to give further doses to correct the problem: but it has become clear that this is not always the case and that significant numbers of patients do go through severe pain, or are at least aware of what is going on, while on the operating table.

There are extremely difficult methodological issues involved in assessing how often this happens. Besides drugs to paralyse patients and remove the pain, they are typically given drugs which erase any memory. Of those who do become conscious during surgery, only a minority report it afterwards. On the other side, many patients may dream or confabulate an awareness that never really existed. Nevertheless the AAGBI's annual surveys are a valuable and creditable effort to provide some data.

The point about the 'third state' does not arise so much from the survey, however, as separate research carried out by Dr Ian F Russell: I haven't been able to track down the particular paper which I think was referred to in Dublin, but there's a useful general article which covers the same experiments here. The techniques used in this case was to keep one forearm unparalysed and then ask the patient at repeated intervals to squeeze with their hand. Russell found that 44% of patients would respond even though they were otherwise apparently unconscious and had no recollection of the episode.

How aware wee these patients? They were able to hear and understand the surgeon and respond appropriately, yet we can, with some caveats, presume that they were not in pain. If they had been in agony, these patients, unlike most, could have waved their hand and would surely have done so - unless the state they were in somehow allowed them to respond to requests but not to initiate any action; or possibly left them too confused even to formulate the simple plan of attracting attention, but not so confused that they could not respond to a simple request. At any rate, it is this curiously ambivalent condition which has prompted Dr Pandit's suggestion of a third state of consciousness.

We might observe that it probably isn't really the third, but perhaps the seventh (or the seventeenth). We already have dreaming, hypnosis, sleepwalking, and meditation to take into account, and very likely we could come up with more. The normal mental state of human beings appears to be complex and layered, and capable of breaking down or degrading in ways you wouldn't have expected.

There's a moral there for artificial general intelligence, I suppose: if you really want to model human thought, you're going to need it to operate on a number of levels at once, rather than having a single 'theatre' or workspace where mental content does its stuff. It may of course be that we don't especially want to model specifically human styles of thought. Perhaps a single level cognitive structure will prove perfectly good for most practical tasks; perhaps it will even be an improvement. That raises the interesting possibility of sitting down to have a conversation with a robot friend who seems to be pretty much like us, but has no unconscious. What would people with no unconscious be like? An outside possibility is that they would be fine at 'easy problem' stuff, but lack phenomenal depth; perhaps they would be the philosophical zombies hose lack of qualia has featured in so many papers.

More generally, the existence of a state in which we remember nothing and experience no pain, but comply helpfully with requests made to us in normal language suggests that either the self as normally envisaged is not as much in executive control as it thinks - a conclusion which would be supported by various other bits of evidence - or that it is less unitary than we generally suppose. Indeed, we might see this as at least persuasive evidence for the existence of a slightly unexpected degree of modularisation. The 'third state' and hypnotic states suggest that there might be a kind of 'implementer' function in the brain. This implementer is responsible for actually getting actions carried out; normally it reacts so fast to the least indication from our self-conscious interior monologue that we don't even notice it; to us it just seems that what we wanted happened. But in certain unusual states the talky, remembering, self-conscious bit of our mind can be turned off while the butler-like implementer is still around and in the absence of the usual boss, quite happy to respond to instructions from outside (and quite capable of using all the language-processing functions of the brain to understand them with).

In its way I find that almost as unsettling as the AAGBI's conclusions about the surprisingly frequent ineffectiveness of anaesthesia.

Are Robots People or People Robots?

29 September 2013

I must admit I generally think of the argument over human-style artificial intelligence as a two-sided fight. There are those who think it's possible, and those who think it isn't. But a chat I had recently made it clear that there are really more differences than that, in particular among those who believe we shall one day have robot chums.

The key difference I have in mind is over whether there really is consciousness at all, or at least whether there's anything special about it.

One school of thought says that there is indeed a special faculty of consciousness; but eventually machines of sufficient complexity will have it too. We may not yet have all the details of how this thing works; maybe we even need some special new secret. But one thing is perfectly clear; there's no magic involved, nothing outside the normal physical account, and in fact nothing that isn't ultimately computable. One day we will be able to build into a machine all the relevant qualities of a human mind. Perhaps we'll do it by producing an actual direct simulation of a human brain, perhaps not; the point is, when we switch on that ultimate robot, it will have feelings and qualia, it will have moral rights and duties, and it will have the same perception of itself as a real existing personality, that we do.

The second school of thought agrees that we shall be able to produce a robot that looks and behaves exactly like a human being. But that robot will not have qualia or feelings or free will or any of the rest of it, because in reality human beings don't have them either! That's one of the truths about ourselves that has been helpfully revealed by the progress of AI: all those things are delusions and always have been. Our feelings that we have a real self, that there is phenomenal experience, and that somehow we have a special kind of agency, those things are just complicated by-products of the way we're organised.

Of course we could split the sceptics too, between those who think that consciousness requires a special spiritual explanation, or is inexplicable altogether, and those who think it is a natural feature of the world, just not computational or not explained by any properties of the physical world known so far. There is clearly some scope for discussion between the former kind of believer and the latter kind of sceptic because they both think that consciousness is a real and interesting feature of the world that needs more explanation, though they differ in their assumptions about how that will turn out. Although there's less scope for discussion, there's also some common ground between the two other groups because both basically believe that the only kind of discussion worth having about consciousness is one that clarifies the reasons it should be taken off the table (whether because it's too much for the human mind or because it isn't worthy of intelligent consideration).

Clearly it's possible to take different views on particular issues. Dennett, for example, thinks qualia are just nonsense and the best possible thing would be to stop even talking about them, while he thinks the ability of human beings to deal with the Frame Problem is a real and interesting ability that robots don't have but could and will once it's clarified sufficiently.

I find it interesting to speculate about which camp Alan Turing would have joined; did he think that humans had a special capacity which computers could one day share, or did he think that the vaunted consciousness of humans turned out to be nothing more than the mechanical computational abilities of his machines? It's not altogether clear, but I suspect he was of the latter school of thought. He notes that the specialness of human beings has never really been proved; and a disbelief in the specialness of consciousness might help explain his caginess about answering the question "can machines think?". He preferred to put the question aside: perhaps that was because he would have preferred to answer; yes, machines can think, but only so long as you realise that 'thinking' is not the magic nonsense you take it to be…

Smonsciousness

20 October 2013

Eric Thomson has some interesting thoughts about aliens, cats, and consciousness, contained in four posts on Brains. In part 1 he proposes a race of aliens who are philosophical zombies - that is, they lack subjective experience; their brains work fine but there is, as it were, no-one home; no qualia. (I have to say that it's not fully clear to me that cats actually have full consciousness in the human sense rather than being fairly single-minded robots seeking food, warmth, etc: but let's not quibble.)

These aliens come to earth and decide, for their own good reasons, to set about quietly studying the brains of cats. They are pretty good at this; they are able to analyse the workings of the cat brain comprehensively and in doing so they discover that feline brains have a special mode of operation. The aliens name this 'smonsciousness'; most of the cats' brain activity is unsmonscious, but certain important functions take place smonsciously. The aliens are able to work out the operation of smonsciousness in some detail and get an understanding of it which is comprehensive except for the minor detail that unbeknownst to them, they don't really get what it actually is. How could they? They have nothing similar themselves.

Part 2 points out that this is a bit of a problem for materialists. The aliens have put together what seems to be a compete materialist account. In one way this seems like a crowning achievement. But it leaves out what it is like for the cats to have experiences. Thomson acknowledges that materialists can claim that this is simply a matter of two different perspectives on the same phenomenon, rather in the way that temperature and mean kinetic energy turn out to be the same thing viewed in different ways. It's a conceptual difference, not an ontological one. But it would be unprecedented to understand something from the lower level without being aware of its higher level version; and in any case, ex hypothesi the aliens know all there is to know about every level of operation of the cat brain. Isn't this an embarrassment for monist materialism?

Part 3 proposes that all might be well if we could wall off phenomenal experience by arguing it needs a special, separate set of concepts which you can only acquire by having phenomenal experience. No amount of playing around with neural concepts will ever get you to phenomenal concepts (rather in the way that Colin McGinn suggests that consciousness is subject to cognitive closure). Neuroscience suffers from semantic poverty in respect of phenomenal experience.

Thomson rightly suggests that this isn't a very comfortable place for the materialist case to rest in and that it would be better for materialists if the semantic poverty idea could be done away with.

So in Part 4 he suggests a cunning manoeuvre. He has actually set the bar for the aliens fairly low: they don't need to have a full grasp of phenomenal experience, hey merely need to become aware that in smonsciousnes something extra is going on (it could be argued that the bar is too

low here, but I think it's OK: if we demand much more we come perilously close to demanding that the aliens actually have phenomenal experience, which is surely too much?).

Now again for their own good reasons the aliens build a simulation of their own consciousness with an additional model which adds smonsciousness when powered up; they call this entity Keanu. Keanu functions fine in alien mode, and when they switch onm his smonsciousness he tells them something is going on which is totally inexpressible, other than by 'Whoa…' Now that may not seem satisfactory, but we can refine it further by supposing the aliens have powerful brain in which they can run the entire simulation. Keanu them is an internal thought experiment: an alien works it through mentally and ends up exclaiming "Dudes! We totally missed phenomenal experience!" Hence the aliens become aware that something is missing and semantic poverty is vanquished.

What do we make of that? There are a lot of angles here; myself I'm suspicious of that interesting concept smonsciousness. It allows the aliens to understand consciousness perfectly in functional terms without knowing what it's really about. Is this plausible? It means that when they consider the following:¶

> *O my Luve's like a red, red rose,*
> *That's newly sprung in June:*
> *O my Luve's like the melodie,*
> *That's sweetly play'd in tune.*

¶….they know exactly and in full what caused Burns to use these words about red things and melodies as he did, but they have no idea at all what the essential significance of the words to Burns was. This is odd. If pressed sufficiently I think we are forced one of two ways. If we cling to the idea that smonsciousness gives a full explanation we are obliged to say that the actual experience of consciousness contributes nothing, and is in fact an epiphenomenon. That's completely tenable, but I would not want to go that way. Alternatively, we have to concede that the aliens don't really have a full understanding, and that smonsciousness isn't really viable. In short, we are more or less back with the dilemma as before.

What about Keanu? Let's consider the internalised version. It's important to remember that running Keanu in your brain does not confer phenomenal experience, it makes you aware that another person of a certain kind would claim phenomenal experience. So what the alien says is not "Dudes! We totally missed phenomenal experience!" , but "Dudes! These cats have a really wild kind of delusion that makes them totally convinced that they're having some kind of special experience - but they can't describe it or what's special about it in any way! What a crazy illusion!". Now Thomson has set the bar low, but surely becoming aware that creatures with smonsciousness claim to be conscious is not quite high enough?

War on Consciousness

3 November 2013

It may be a little off our usual beat, but Graham Hancock's piece in the New Statesman raised some interesting thoughts.

It's the 'war on drugs' that is Hancock's real target, but he says it's really a war on consciousness...¶

> *This extraordinary imposition on adult cognitive liberty is justified by the idea that our brain activity, disturbed by drugs, will adversely impact our behaviour towards others. Yet anyone who pauses to think seriously for even a moment must realize that we already have adequate laws that govern adverse behaviour towards others and that the real purpose of the "war on drugs" must therefore be to bear down on consciousness itself.*

That doesn't seem quite right. It's true there are weak arguments for laws against drugs - some of them based on bad consequences that arguably arise from the laws rather than the drugs - but there are reasonable ones, too. The bedrock point is that taking a lot of psychoactive drugs is probably bad for you. Hancock and many others might say that we should have the right to harm ourselves, or at any rate to risk harm, if we don't hurt anyone else, but that principle is not, I think, generally accepted by most legislatures. Moreover there are special arguments in the case of drugs. One is that they are addictive. 'Addiction' is used pretty widely these days to cover any kind of dependency or habit, but I believe the original meaning was that an addict became physically dependent, unable to stop taking the drug without serious, possibly even fatal consequences, while at the same time ever larger doses were needed to achieve relief. That is clearly not a good way to go, and it's a case where leaving people to make up their own minds doesn't really work because of the dependency. Secondly, drugs may affect the user's judgement and for that reason too should arguably be a case where people are not left to judge risks for themselves.

Now, as a matter of fact neither of those arguments may apply in the case of some restricted drugs - they may not be addictive in that strongest sense and they may not remove the user's ability to judge risks; and the risks themselves may in some cases have been overstated; but we don't have to assume that governments are simply set on denying us the benefits of enhanced consciousness.

What would those benefits be? They might be knowledge, enhanced cognition, or simple pleasure. We could also reverse the argument that Hancock attributes to our rulers and suggest that drugs make people less likely to harm others. People who are lying around admiring the wallpaper in a confused manner are not out committing crimes, after all.

Enhanced cognition might work up to a point in some cases: certain drugs really do help dispel fatigue or anxiety and sharpen concentration in the short term. But the really interesting possibility for us is that drug use might allow different kinds of cognition and knowledge. I think the evidence on fathoming the secrets of the Universe is rather discouraging. Drugs may often make people feel as if they understand everything, but it never seems to be possible to write the

insights down. Where they are written down, they turn out to be like the secret of the cosmos apprehended by Oliver Wendell Holmes under the influence of ether; later he discovered his notes read “A strong smell of turpentine prevails throughout”.

But perhaps we’re not dealing with that kind of knowledge. Perhaps instead drugs can offer us the kind of ineffable knowledge we get from qualia? Mary the colour scientist is said to know something new once she has seen red for the first time; not something about colour that could have been written down, or ex hypothesi she would have known it already, but what it is like. Perhaps drugs allow us to experience more qualia, or even super qualia; to know what things are like whose existence we should not otherwise have suspected. Terry Pratchett introduced the word ‘knurd’ to describe the state of being below zero on the drunkenness scale; needing a drink to bring you up to the normal mental condition: perhaps philosophical zombies, who experience no qualia, are simply in a similar state with respect to certain drugs.

That seems plausible enough, but it raises the implication that normal qualia are also in fact delusions (not an uncongenial implication for some). For drugs there is a wider problem of non-veridicality. We know that drugs can cause hallucinations, and as mentioned above, can impart feelings of understanding without the substance. What if it’s all like that? What if drug experiences are systematically false? What if we don’t really have any new knowledge or any new experiences on drugs, we just feel as if we have? For that matter, what about pleasure? What if drugs give us a false memory of having had a good time - or what if they make us think we’re having a good time now although in reality we’re not enjoying it at all? You may well feel that last one is impossible, but it doesn’t pay to underestimate the tricksiness of the mind.

Well, many people would say that the feeling of having had a good time is itself worth having, even if the factual element of the feeling is false. So perhaps in the same way we can say that even if qualia are delusions, they’re valuable ones. Perhaps the exalted places to which drugs take us are imaginary; but just because somewhere doesn’t exist doesn’t mean it isn’t worth going there. For myself I generally prefer the truth (no argument for that, just a preference) and I think I generally get it most reliably when sober and undrugged.

Hancock, at any rate, has another kind of knowledge in mind. He suggests that the brain may turn out to be, not a generator of consciousness but rather a receiver, tuned in to the psychic waves where, I assume, our spiritual existence is sustained. Drugs, he proposes, might possibly allow us to twiddle the knobs on our mental apparatus so as to receive messages from others: different kinds of being or perhaps people in other dimensions. I’m not quite clear where he draws the line between receiving and existing, or whether we should take ourselves to be in the brain or in the spiritual ether. If we’re in the brain, then the signals we’re receiving are a form of outside control which doesn’t sound very nice: but if we’re really in the ether then when the signals from other beings are being received by the brain we ought to lose consciousness, or at least lose control of our bodies, not just pick up a message. No doubt Hancock could clarify, given a chance, but it looks as if there’s a bit of work to be done.

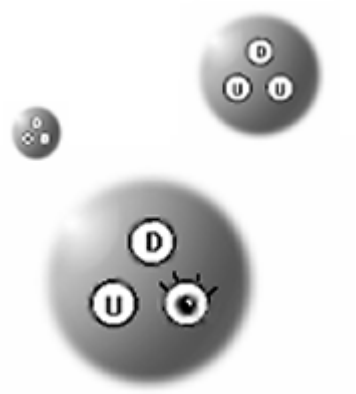
But let’s not worry too much, because the idea of the brain as a mere receiver seems highly dubious. We know now that very detailed neuronal activity is associated with very specific mental content, and as time goes on that association becomes ever sharper. This means that if the brain is a receiver the signals it receives must be capable of influencing a vast collection of close-packed neurons in incredibly exquisite detail. It’s only fair to remember that a neurologist as distinguished as Sir John Eccles, not all that long ago, thought this was exactly what was happening; but to me it seems incompatible with ordinary physics. We can manipulate small

areas of the brain from outside with suitable equipment, but dictating its operation at this level of detail, and without any evident physical intervention seems too much. Hancock says the possibility has not been disproved, and for certain standards of proof that's right; but I reckon by the provisional standards that normally apply for science we can rule out the receiver thesis.

Speaking of manipulating the brain from outside, it seems inevitable to me that within a few years we shall have external electronic means of simulating the effects of certain drugs, or at least of deranging normal mental operation in a diverting and agreeable way. You'll be able to slip on a helmet, flick a switch, and mess with your mind in all sorts of ways. They might call it e-drugs or something similar, but you'll no longer need to buy dodgy chemicals at an exorbitant mark-up. What price the war on drugs or on consciousness then?

Quark Consciousness

18 November 2013



One of the main objections to panpsychism, the belief that mind, or at any rate experience, is everywhere, is that it doesn't help. The point of a theory is to take an issue that was mysterious to begin with and make it clear; but panpsychism seems to leave us with just as much explaining to do as before. In fact, things may be worse. To begin with we only needed to explain the occurrence of consciousness in the human brain; once we embrace panpsychism we have to explain it's occurrence everywhere and account for the difference between the consciousness in a lump of turf and the consciousness in our heads. The only way that could be an attractive option would be if there were really good and convincing answers to these problems ready to hand.

Creditably, Patrick Lewtas recognises this and rolling up his sleeves has undertaken the job of explaining first, how 'basic bottom level experience' makes sense, and second, how it builds up to the high-level kind of experience going on in the brain. A first paper, tackling the first question, "What is it like to be a Quark" appeared in the JCS recently (Alas, there doesn't seem to be an online version available to non-subscribers.)

Lewtas adopts an idiosyncratic style of argument, loading himself with Constraints like a philosophical Houdini.¶

1. Panpsychism should attribute to basic physical objects all but only those types of experiences needed to explain higher-level (including, but not limited to, human) consciousness.
2. Panpsychism must eschew explanatory gaps.
3. Panpsychism must eschew property emergence.
4. Maximum possible complexity of experience varies with complexity of physical structure.
5. Basic physical objects have maximally simple structures. They lack parts, internal structure, and internal processes.
6. Where possible and appropriate, panpsychism should posit strictly-basic conscious properties similar, in their higher-order features to strictly-basic physical properties.
7. Basic objects with strictly-basic experiences have the constantly and continuously.
8. Each basic experience-type, through its strictly-basic instances. characterizes (at least some) basic physical objects.

Of course it is these very constraints that end up getting him where he wanted to be all along. To justify each of them and give the implications would amount to reproducing the paper; I'll try to summarise in a freer style here.

Lewtas wants his basic experience to sit with basic physical entities and he wants it to be recognisably the same kind of thing as the higher level experience. This parsimony is designed to avoid any need for emergence or other difficulties; if we end up going down that sort of road, Lewtas feels we will fall back into the position where our theory is too complex to be attractive in competition with more mainstream ideas. Without seeming to be strongly wedded to them, he

chooses to focus on quarks as his basic unit, but he does not say much about the particular quirks of quarks; he seems to have chosen them because they may have the property he's really after; that of having no parts.

The thing with no parts! Aiee! This ancient concept has stalked philosophy for thousands of years under different names: the atom, a substance, a monad (the first two names long since passed on to other, blameless ideas). I hesitate to say that there's something fundamentally problematic with the concept itself (it seems to work fine in geometry); but in philosophy it seems hard to handle without generating a splendid effusion of florid metaphysics. The idea of yoking it together with the metaphysically tricky modern concept of quarks makes my hair stand on end. But perhaps Lewtas can keep the monster in check: he wants it, presumably, because he wants to build on bedrock, with no question of basic experience being capable of further analysis.

Some theorists, Lewtas notes, have argued that the basic level experience of particles must be incomprehensible to us; as incomprehensible as the experiences of bats according to Nagel, or indeed even worse. Lewtas thinks things can, and indeed must, be far simpler and more transparent than that. The experience of a quark, he suggests, might just be like the simple experience of red; red detached from any object or pattern, with no limits or overtones or significance; just red. Human beings can most probably never achieve such simplicity in its pure form, but we can move in that direction and we can get our heads around 'what it's like' without undue difficulty.

Now the partless thing begins to give trouble; a thing which has no parts cannot change, because change would imply some kind of reorganisation or substitution; you can't rearrange something that has no parts and if you substitute anything you have to substitute another whole thing for the first one, which is not change but replacement. At best the thing's external relations can change. If one of the properties of the quark is an experience of red, therefore, that's how it stays. It carries on being an experience of red, and it does not respond in any way to its environment or anything outside itself. I think we can be forgiven if we already start to worry a little about how this is going to work with a perceptual system, but that is for the later paper.

Lewtas is aware that he could be in for an awfully large catalogue of experiences here if every possible basic experience has to be assigned to a quark. His hope is that some experiences will turn out to be composites, so that we'll be able to make do with a more restricted set: and he gives the example of orange experience reducing to red and yellow experience. A bad example: orange experience is just orange experience, actually, and the fact that orange paint can be made by mixing red and yellow paint is just a quirk of the human visual system, not an essential quality of orange light or orange phenomenology. A bad example doesn't mean the thesis is false; but a comprehensive reduction of phenomenology to a manageable set of basic elements is a pretty non-trivial requirement. I think in fact Lewtas might eventually be forced to accept that he has to deal with an infinite set of possible basic experiences. Think of the experience of unity, duality, trinity… That's debatable, perhaps.

At any rate Lewtas is prepared to some extent. He accepts explicitly that the number of basic experiences will be greater than the number of different kinds of basic quark, so it follows that basic physical units must be able to accommodate more than one basic experience at the same time. So your quark is having a simple, constant experience of red and at the same time it's having a simple, constant experience of yellow.

That has got to be a hard idea for Lewtas to sell. It seems to risk the simple transparency which was one of his main goals, because it is surely impossible to imagine what having two or more completely pure but completely separate experiences at the same time is like. However, if that bullet is bitten, then I see no particular reason why Lewtas shouldn't allow his quarks to have all possible experiences simultaneously (my idea, not his).

By the time we get to this point I find myself wondering what the quarks, or the basic physical units, are contributing to the theory. It's not altogether clear how the experiences are anchored to the quarks and since all experiences are going to have to be readily available everywhere, I wonder whether it wouldn't simplify matters to just say that all experiences are accessible to all matter. That might be one of the many issues cleared up in the paper to follow where perhaps, with one cat-like leap, Lewtas will escape the problems which seem to me to be on the point of having him cornered.

More Philosophy?

1 December 2013

Why is it that we can't solve the mind/body problem? Well, if we define that problem as being about the capacity of mental events to cause physical events, there is a project in progress at Durham University that says it's about the lack of good philosophy, or more specifically, that our problem stems from inadequate ontology (ontology being the branch of metaphysics that deals with what there is). The project has been running for a few years, and now a substantial volume of corrective metaphysics has been published, with a thoughtful and full review here. (Hat-tip to Micha for drawing this to my attention).

The book is not a manifesto, because the authors do not share a single view: it's more like an exhibition. What's on offer here is a variety of philosophical views of mental causation, all more sophisticated than the ones we typically encounter in discussions of artificial intelligence. The review gives a good sense of what's on offer, and depending on your inclinations you may see it as a collection of esoteric and unhelpful complications, or as a magic sweetshop whose every jar holds the way to a new world of possibility and enlightenment. I think the average view will see it as a bookshop with many volumes of dull sermons and outdated almanacs which might nevertheless just be holding somewhere in the dusty back room that one book that makes sense of everything.

Is it likely that better philosophy will deliver the answer? There is a horrid vision in my mind in which the neurologists and/or the AI people produce a model which seems to work; we're able to build machines which talk to us in the same way as human beings, and we can explain exactly how the brain does its stuff and how it is analogous to these machines: but the philosophers go on doubting whether this machine consciousness is real consciousness. No-one else cares.

Moreover, there are some identifiable weaknesses in philosophy which are clearly on display in the current volume. First is the fissiparous nature of philosophical discussion. I said this was an exhibition rather than a manifesto; but wouldn't a manifesto have been better? It's not achievable because every philosopher has his or her own view and the longer discussion goes on the more possible views there are. In one way it's a pleasing, exploratory quality, but if you want a solution it's a grave handicap. Second, and related, there's no objective test beyond logical consistency. Experiments will never prove any of these views wrong.

Third, although philosophy is too difficult, it's also too easy. Someone somewhere once said that Aristotle's problem was that he was too clever. For him, it was always possible to come up with a theory which justified the outlook of a complacent middle-aged Ancient Greek: theories which have turned out, so far as we can test them, to be almost invariably false or incomplete. Less clever pre-Socratic philosophers like Heraclitus or Parmenides were forced to adopt weirder points of view which in the long run might actually tell us more.

The current volume, I think, might contain many cases of clever people making cases that broadly justify common sense while the real truth may be out there in the wild regions beyond. E.J.Lowe, one of the editors of the book and champions of the project, has a view about the powers of the will. He characterises powers as active or passive on the one hand and causal or non-causal. This leaves open the possibility of a power which is both active and non-causal. He wants the human will to have these properties, so that it is spontaneous and not causally inefficacious without the agent per se thereby bringing about any sort of effect (if I've got that right). The spontaneity is supposed to resemble the spontaneity of the decay of a specific radium atom, and hence be consistent with physics, while the causal efficacy is of a kind that does not require an interruption of normal physics while still being an important corollary of our status as rational beings.

This is clever stuff, no doubt, but it looks like an attempt - you may consider it a doomed attempt - to explain away the problems with our common sense views rather than correcting them. We're being offered loopholes which may - debatably - let us carry on thinking what we've always thought, rather than offering us a new perspective. It leaves me feeling the way I might feel after a clever lawyer has explained why his client should not be convicted; yeah, but did he do it? There's no 'aha!' moment on offer. In her review Sara Bernstein suggests that sceptics may be inclined to turn back to reductionism, and I must confess that is indeed my inclination.

Still, I can't shake my hope that somewhere in that dusty old bookshop the truth is to be found, and so I can't help wishing the project well.

The Will to Persist

15 December 2013

I was interested to see reports here and there the other day that scientists had discovered the seat of the will in the anterior midcingulate cortex.

That's not precisely the case, of course; there's an article here which describes the research. The scientists in question had an unusual opportunity to use electrodes in the brains of two patients; although they did indeed operate in the anterior midcingulate cortex they believe they were stimulating the brain's salience network, which is quite widely distributed. The effect was apparently to create feelings of needing to persist against challenging circumstances; the researchers themselves call it "the will to persevere". The patients were fully conscious and able to describe their feelings, but alas no tests were carried out to see whether they were in fact more persistent when stimulated.

The correct interpretation of the results seems difficult to me. As I understand it, the theory of the salience network holds that brain activity is controlled by neural networks which stretch across several regions of the brain. The default mode network, or DMN, is the one that operates when we're not focused on anything in particular, perhaps daydreaming. It has been suggested that loss of this function is what distinguishes people who have "locked-in" syndrome from those who are in a "persistent vegetative state" - if you lose your DMN you're not really there any more, in other words.

When we concentrate on a task, another network takes over - the central executive network, or CEN. The role of the salience network, if I've got this right, is primarily to act as arbitrator between the two. It spots something that deserves attention - something salient, indeed - and switches control from DMN to CEN. That's fine, but it doesn't seem to have much to do with persistence; it's actually about changing the object of attention, not sticking with it. But perhaps strong or continuing stimulation of the salience network has that kind of effect. The salience network says "you need to look at this", so perhaps when it operates emphatically we think "yes, I'm going to look the hell out of that alright; I'm going to look at that intensively; when I've finished looking at that, by golly it's going to stay looked at".

More plausibly it might all be to do with physiological effects; besides directing attention the salience network has a role in gearing up our "fight or flight" state, and it might just be that in that state we feel ready for a challenge (in which case a readiness to persist comes into it, but surely isn't the whole point); that would be a William James style emotion, originally a matter of the gut more than the mind.

Anyway, this really has nothing to do with an organ of the will. That is an interesting notion, though, isn't it? My own assumption is that the will emerges from the operation of general cognition and that there couldn't be a separate will module. If such a module determined the actions to be willed, it would surely have to encompass almost the whole of cognition, and so be far more than just a module; if it merely willed the actions selected for it by the rest of the brain it wouldn't amount to much at all.

People do, of course, often hypothesise that there might be a special function for assigning value or flagging up those things we ought to pursue as desirable. To me, though, that seems a bit different; the will, properly understood, is not a matter of basic motivation, but a faculty which might over-ride that motivation, either by operating at a meta level or simply by acting as a restraining and countervailing force.

Would that even be a distinct faculty of its own? Some would probably question whether talking about the will is a useful approach at all, rather than a relic from outmoded ideas about the soul controlling the body through acts of will. I must admit I find it hard to think of any subject that can't be adequately discussed without mentioning the will. Even free will doesn't really lose anything if we talk about free action. So is the will even worth persisting with? I can feel my DMN kicking in.

Consciousness and Alcohol

22 December 2013

It has always seemed remarkable to me that the ingestion of a single substance can have such complex effects on behaviour. Alcohol does it, in part, by promoting the effects of inhibitory neurotransmitters and suppressing the effects of excitatory ones, while also whacking up a nice surge of dopamine - or so I understand. This messes up co-ordination and can lead to loss of memory and indeed consciousness; but the most interesting effect, and the one for which alcohol is sometimes valued, is that it causes disinhibition. This allows us to relax and have a good time but may also encourage risky behaviour and lead to us saying things - in vino veritas - we wouldn't normally let out.

Curiously, though, there's no solid scientific support for the idea that alcohol causes disinhibition, and good evidence that alcohol is blamed for disinhibition it did not cause. One of the slippery things about the demon drink is that its effects are strongly conditioned by the drinkers expectations. It has been shown that people who merely thought they were consuming alcohol were disinhibited just as if they had been; while other studies have shown that risky sexual behaviour can actually be deterred in those who have had a few drinks, if the circumstances are right.

One piece of research suggests that meta-consciousness is impaired by alcohol; drink makes us less aware of our own mental state. But a popular and well-supported theory these days is that drinking causes 'alcohol myopia'. On this theory, when we're drunk we lose track of long-term and remote factors, while our immediate surroundings seem more salient. One useful aspect of the theory is that it explains the variability of the effects of alcohol. It may make remoter worries recede and so leave us feeling unjustifiably happy with ourselves; but if reminders of our problems are close while the long term looks more hopeful, the effect may be depressing. Apparently subjects who had the words 'AIDS KILLS' actually written on their arm were less likely to indulge in risky sex (I suspect it might kind of dent your chances of getting a casual partner, actually).

Not a Panpsychist But An Emergentist?

5 January 2014

Christof Koch declares himself a panpsychist in this interesting piece, but I don't think he really is one. He subscribes to the Integrated Information Theory (IIT) of Giulio Tononi, which holds that consciousness is created by the appropriate integration of sufficient quantities of information. The level of integrated information can be mathematically expressed in a value called Phi: we have discussed this before a couple of times. I think this makes Koch an emergentist, but curiously enough he vigorously denies that.

Koch starts with a quotation about every outside having an inside which aptly brings out the importance of the first-person perspective in all these issues. It's an implicit theme of what Koch says (in my reading at least) that consciousness is something extra. If we look at the issue from a purely third-person point of view, there doesn't seem to be much to get excited about. Organisms exhibit different levels of complexity in their behaviour and it turns out that this complexity of behaviour arises from a greater complexity in the brain. You don't say! The astonishment meter is still indicating zero. It's only when we add in the belief that at some stage the inward light of consciousness, actual phenomenal experience, has come on that it gets interesting. It may be that Koch wants to incorporate panpsychism into his outlook to help provide that ineffable light, but attempting to make two theories work together is a risky path to take. I don't want to accuse anyone of leaning towards dualism (which is the worst kind of philosophical bitchiness) but... well, enough said. I think Koch would do better to stick with the austere simplicity of IIT and say: that magic light you think you see is just integrated information. It may look a bit funny but that's all it is, get used to it.

He starts off by arguing persuasively that consciousness is not the unique prerogative of human beings. Some, he says, have suggested that language is the dividing line, but surely some animals, preverbal infants and so on should not be denied consciousness? Well, no, but language might be interesting, not for itself but because it is an auxiliary effect of a fundamental change in brain organisation, one that facilitates the handling of abstract concepts, say (or one that allows the integration of much larger quantities of information, why not?). It might almost be a side benefit, but also a handy sign that this underlying reorganisation is in place, which would not be to say that you couldn't have the reorganisation without having actual language. We would then have something, human-style thought, which was significantly different from the feelings of dogs, although the impoverishment of our vocabulary makes us call them both consciousness.

Still, in general the view that we're dealing with a spectrum of experience, one which may well extend down to the presumably dim adumbrations of worms and insects, seems only sensible.

One appealing way of staying monist but allowing for the light of phenomenal experience is through emergence: at a certain level we find that the whole becomes more than the sum of its

parts: we do sort of get something extra, but in an unobjectionable way. Strangely, Koch will have no truck with this kind of thinking. He says:

> *'the mental is too radically different for it to arise gradually from the physical'.*

At first sight this seemed to me almost a direct contradiction of what he had just finished saying. The spectrum of consciousness suggests that we start with the blazing 3D cinema projector of the human mind, work our way down to the magic lanterns of dogs, the candles of newts, and the faint tiny glows of worms – and then the complete darkness of rocks and air. That suggests that consciousness does indeed build up gradually out of nothing, doesn't it? An actual panpsychist, moreover, pushes the whole thing further, so that trees have faint twinkles and even tiny pieces of clay have a detectable scintilla.

Koch's view is not, in fact, contradictory: what he seems to want is something like one of those dimmer switches that has a definite on and off, but gradations of brightness when on. He's entitled to take that view, but I don't think I agree that gradual emergence of consciousness is unimaginable. Take the analogy of a novel. We can start with Pride and Prejudice, work our way down through short stories or incoherent first drafts, to recipe books or collections of limericks, books with scribble and broken sentences, down to books filled with meaningless lines, and the chance pattern of cracks on a wall. All the way along there will be debatable cases, and contrarians who disbelieve in the real existence of literature can argue against the whole thing ('You need to exercise your imagination to make Pride and Prejudice a novel; but if you are willing to use your imagination I can tell you there are finer novels in the cracks on my wall than anything Jane bloody Austen ever wrote…') : but it seems clear enough to me that we can have a spectrum all the way down to nothing. That doesn't prove that consciousness is like that, but makes it hard to assert that it couldn't be.¶The other reason it seems odd to hear such an argument from Koch is that he espouses the IIT which seems to require a spectrum which sits well with emergentism. Presumably on Koch's view a small amount of integrated information does nothing, but at some point, when there's enough being integrated, we start to get consciousness? Yet he says:

> *"if there is nothing there in the first place, adding a little bit more won't make something. If a small brain won't be able to feel pain, why should a large brain be able to feel the god-awfulness of a throbbing toothache? Why should adding some neurons give rise to this ineffable feeling?"*

Well, because a small brain only integrates a small amount of information, whereas a large on integrates enough for full consciousness? I think I must be missing something here, but look at this.

> *" [Consciousness] is a property of complex entities and cannot be further reduced to the action of more elementary properties. We have reached the ground floor of reductionism."*

Isn't that emergence? Koch must see something else which he thinks is essential to emergentism which he doesn't like, but I'm not seeing it.

The problem with Koch being panpsychist is that for panpsychists souls (or in this case consciousness) have to be everywhere. Even a particle of stone or a screwed-up sheet of wrapping paper must have just the basic spark; the lights must be at least slightly on. Koch doesn't want to go quite that far - and I have every sympathy with that, but it means taking the

pan out of the panpsychist. Koch fully recognises that he isn't espousing traditional full-blooded panpsychism but in my opinion he deviates too far to be entitled to the badge. What Koch believes is that everything has the potential to instantiate consciousness when correctly organised and integrated. That amounts to no more than believing in the neutrality of the substrate, that neurons are not essential and that consciousness can be built with anything so long as its functional properties are right. All functionalists and a lot of other people (not everyone, of course) believe that without being panpsychists.

Perhaps functionalism is really the direction Koch's theories lean towards. After all, it's not enough to integrate information in any superficial way. A big database which exhaustively cross-referenced the Library of Congress would not seem much of a candidate for consciousness. Koch realises that there have to be some rules about what kinds of integration matter, but I think that if the theory develops far enough these other constraints will play an increasingly large role, until eventually we find that they have taken over the theory and the quantity of integrated information has receded to the status of a necessary but not sufficient condition.

I suppose that that might still leave room for Tononi's Phi meter, now apparently built, to work satisfactorily. I hope it does, because it would be pretty useful.

Do Zombies Have Rights?

19 January 2014

Existential Comics raised an interesting question (thanks to Micha for pointing it out). In the strip a doctor with a machine that measures consciousness (rather like Tononi's new machine, except that that measures awareness) tells an unlucky patient he lacks the consciousness-producing part of the brain altogether. Consequently, the doctor says, he is legally allowed to harvest the patient's organs.

Would that be right?

We can take it that what the patient lacks is consciousness in the 'Hard Problem' sense. He can talk and behave quite normally, it's just that when he experiences things there isn't 'something it is like'; there's no real phenomenal experience. In fact, he is a philosophical zombie, and for the sake of clarity I take him to be a strict zombie; one of the kind who are absolutely like their normal human equivalent in every important detail except for lacking qualia (the cartoon sort of suggests otherwise, since it implies an actual part of the brain is missing, but I'm going to ignore that).

Would lack of qualia mean you also lacked human rights and could be treated like an animal, or worse? It seems to me that while lack of qualia might affect your standing as a moral object (because it would bear on whether you could suffer, for example), it wouldn't stop you being a full-fledged moral subject (you would still have agency). I think I would consequently draw a distinction between the legal and the moral answer. Legally, I can't see any reason why the absence of qualia would make any difference. Legal personhood, rights and duties are all about actions and behaviour, which takes us squarely into the realm of the Easy Problem. Our zombie friend is just like us in these respects; there's no reason why he can't enter into contracts, suffer punishments, or take on responsibilities. The law is a public matter; it is forensic - it deals with the things dealt with in the public forum; and it follows that it has nothing to say about the incorrigibly private matter of qualia.

Of course the doctor's machine changes all that and makes qualia potentially a public matter (which is one reason why we might think the machine is inherently absurd, since public qualia are almost a contradiction in terms). It could be that the doctor is appealing to some new, recently-agreed legislation which explicitly takes account of his equipment and its powers. If so, such legislation would presumably have to have been based on moral arguments, so whichever way we look at it, it is to the moral discussion that we must turn.

This is a good deal more complicated. Why would we suppose that phenomenal experience has moral significance? There is a general difficulty because the zombie has experiences too. In conditions when a normal human would feel fear, he trembles and turns pale; he smiles and relaxes under the influence of pleasure; he registers everything that we all register. He writes love poetry and tells us convincingly about his feelings and tastes. It's just that, on the inside, everything is hollow and void. But because real phenomenal experience always goes along with zombie-style experience, it's hard for us to find any evidence as to why one matters when the other doesn't.

The question also depends critically on what ethical theories we adopt. We might well take the view that our existing moral framework is definitive, authorised by God or tradition, and therefore if it says nothing about qualia, we should take no account of them either. No new laws are necessary, and there can be no moral reason to allow the harvesting of organs.

In this respect I believe it is the case that medieval legislatures typically saw themselves, not as making new laws, but as rediscovering the full version of old ones, or following out the implications of existing laws for new circumstances. So when the English parliamentarians wanted to argue against Charles I's Ship Tax, rather than rest their case on inequity, fiscal distortion, or political impropriety, they appealed to a dusty charter of Ine, ancient ruler of Wessex (regrettably they referred to Queen Ine, whereas he had in fact been a robustly virile King).

Even within a traditional moral framework, therefore, we might find some room to argue that new circumstances called for some clarification; but I think we would find it hard going to argue for the harvesting.

What if we were utilitarians, those people who say that morality is acting to bring about the greatest happiness of the greatest number? Here we have a very different problem because the utilitarians are more than happy to harvest your organs anyway if by doing so they can save more than one person, no matter whether you have qualia or not. This unattractive kind of behaviour is why most people who espouse a broadly utilitarian framework build in some qualifications (they might say that while organ harvesting is good in principle actual human aversion to it would mean that in practice it did not conduce to happiness overall, for example).

The interesting point is whether zombie happiness counts towards the utilitarian calculation. Some might take the view that without qualia it had no real value, so that the zombie's happiness figure should be taken as zero. Unfortunately there is no obvious answer here; it just depends what kind of happiness you think is important. In fact some consequentialists take the utilitarian system but plug into it desiderata other than happiness anyway. It can be argued that old-fashioned happiness utilitarianism would lead to us all sitting in boxes that directly stimulated our pleasure centres, so something more abstract seems to be needed; some even just speak of 'utility' without making it any more specific.

No clear answer then, but it looks as if qualia might at least be relevant to a utilitarian.

What about the Kantians? Kant, to simplify somewhat, thought we should act in accordance with the kind of moral rules we should want other people to adopt. So, we should be right to harvest the organs so long as we were content that if we ourselves turned out to be zombies, the same thing would happen to us. Now I can imagine that some people might attach such value to qualia that they might convince themselves they should agree to this proposition; but in general the answer is surely negative. We know that zombies behave exactly like ordinary people, so they would not for the most part agree to having their organs harvested; so we can say with confidence that if I were a zombie I should still tell the doctor to desist.

I think that's about as far as I can reasonably take the moral survey within the scope of a blog post. At the end of the day, are qualia morally relevant? People certainly talk as if they are in some way fundamental to value. "Qualia are what make my life worth living" they say: unfortunately we know that zombies would say exactly the same.

I think most people, deliberately or otherwise, will simply not draw a distinction between real phenomenal experience on one hand and the objective experience of the kind a zombie can have on the other. Our view of the case will in fact be determined by what we think about people with and without feelings of both kinds, rather than people with and without qualia specifically. If so, qualia sceptics may find that grist to their mill.

Micha has made some interesting comments which I hope he won't mind me reproducing.

> *The question of deontology vs consequentialism might be involved. A deontologist has less reason -- although still some -- to care about the content of the victim's mind. Animals are also objects of morality; so the whole question may be quantitative, not qualitative.*
>
> *Subjects like ethics aren't easy for me to discuss philosophically to someone of another faith. Orthodox Judaism, like traditional Islam, is a legally structured religion. Therefore ethics aren't discussed in the same language as in western society, since how the legal system processes revelation impacts conclusion.*
>
> *In this case, it seems relevant that the talmud says that someone who kills adnei-hasadeh (literally: men of the field) is as guilty of murder as someone who kills a human being. It's unclear what the talmud is referring to: it may be a roman mythical being who is like a human, but with an umbilicus that grows down to roots into the earth, or perhaps an orangutan -- from the Malay for "man of the jungle", or some other ape. Whatever it is, only actual human beings are presumed to have free will. And yet killing one qualifies as murder, not the killing of an animal.*

The Wonder of Consciousness

1 February 2014

Harold Langsam's new book is a bold attempt to put philosophy of mind back on track. For too long, he declares, we have been distracted by the challenge from reductive physicalism. Its dominance means that those who disagree have spent all their time making arguments against it, instead of developing and exploring their own theories of mind. The solution is that, to a degree, we should ignore the physicalist case and simply go our own way. Of course, as he notes, setting out a rich and attractive non-reductionist theory will incidentally strengthen the case against physicalism. I can sympathise with all that, though I suspect the scarcity of non-reductive theorising also stems in part from its sheer difficulty; it's much easier to find flaws in the reductionist agenda than to develop something positive of your own.

So Langsam has implicitly promised us a feast of original insights; what he certainly gives us is a bold sweep of old-fashioned philosophy. It's going to be a priori all the way, he makes clear; philosophy is about the things we can work out just by thinking. In fact a key concept for Langsam is intelligibility; by that, he means knowable a priori. It's a usage far divorced from the normal meaning; in Langsam's sense most of the world (and all books) would be unintelligible.

The first target is phenomenal experience; here Langsam is content to use the standard terminology although for him phenomenal properties belong to the subject, not the experience. He speaks approvingly of Nagel's much-quoted formulation 'there is something it is like' to have phenomenal experience, although I take it that in Langsam's view the 'it' that something is like is the person having the experience, which I don't think was what Nagel had in mind. Interestingly enough, this unusual feature of Langsam's theory does not seem to matter as much as we might have expected. For Langsam, phenomenal properties are acquired by entry into consciousness, which is fine as far as it goes, but seems more like a re-description than an explanation.

Langsam believes, as one would expect, that phenomenal experience has an inexpressible intrinsic nature. While simple physical sensations have structural properties, in particular, phenomenal experience does not. This does not seem to bother him much, though many would regard it as the central mystery. He thinks, however, that the sensory part of an experience - the unproblematic physical registration of something - and the phenomenal part are intelligibly linked. In fact, the properties of the sensory experience determine those of the phenomenal experience. In sensory terms, we can see that red is more similar to orange than to blue, and for Langsam it follows that the phenomenal experience of red similarly has an intelligible similarity to the phenomenal experience of orange. In fact, the sensory properties explain the phenomenal ones.

This seems problematic. If the linkage is that close, then we can in fact describe phenomenal experience quite well; it's intelligibly like sensory experience. Mary the colour scientist, who has never seen colours, actually will not learn anything new when she sees red: she will just confirm that the phenomenal experience is intelligibly like the sensory experience she already understood perfectly. In fact because the resemblance is intelligible - knowable a priori - she could work out what it was like before seeing red at all. To that Langsam might perhaps reply

that by ‘a priori’ he means not just pure reasoning but introspection, a kind of internal empiricism.

It still leaves me with the feeling that Langsam has opened up a large avenue for naturalisation of phenomenal experience, or even suggested that it is in effect naturalised already. He says that the relationship between the phenomenal and the sensory is like the relation between part and whole; awfully tempting, then, to conclude that his version of phenomenal experience is merely an aspect of sensory experience, and that he is much more of a sceptic about phenomenality than he realises.

This feeling is reinforced when we move on to the causal aspects. Langsam wants phenomenal experience to have a role in making sensory perceptions available to attention, through entering consciousness. Surely this is making all the wrong people, from Langsam’s point of view, nod their heads: it sounds worryingly functionalist. Langsam wants there to be two kinds of causation: ‘brute causation’, the ordinary kind we all believe in, and intelligible causation, where we can just see the causal relationship. I enjoyed Langsam taking a pop at Hume, who of course denied there was any such thing; he suggests that Hume’s case is incomplete, and actually misses the most important bits. In Langsam’s view, as I read it, we just see inferences, perceiving intelligible relationships.

The desire to have phenomenal experience play this role seems to me to carry Langsam too far in another respect: he also claims that simply believing that p has a phenomenal aspect. I take it he wishes this to be the case so that this belief can also be brought to conscious attention by its phenomenal properties, but look; it just isn’t true. ‘Believing that p’ has no phenomenal properties whatever; there is nothing it is like to believe that p, in the way that there is something it is like to see a red flower. The fact that Langsam can believe otherwise reinforces the sense that he isn’t such a believer in full-blooded phenomenality as he supposes.

We can’t accuse him of lacking boldness, though. In the second part of the book he goes on to consider appropriateness and rationality; beliefs can be appropriate and rational, so why not desires? At this point we’re still apparently engaged on an enquiry into philosophy of mind, but in fact we’ve also started doing ethics. In fact I don’t think it’s too much of a stretch to say that Langsam is after Kant’s categorical imperative. Our desires can stem intelligibly from such sensations as pain and pleasure, and our attitudes can be rational in relation to the achievement of desires. But can there be globally rational desires - ones that are rational whatever we may otherwise want?

Langsam’s view is that we perceive value in things indirectly through our feelings and when our desires are for good things they are globally rational. If we started out with Kant, we seem to have ended up with a conclusion more congenial to G.E,Moore. I admire the boldness of these moves, and Langsam fleshes out his theory extensively along the way - which may be the real point as far as he’s concerned. However, there are obvious problems about rooting global rationality in something as subjective and variable as feelings, and without some general theory of value Langsam’s system is bound to suffer a certain one-leggedness.

I do admire the overall boldness and ambition of Langsam’s account, and it is set out carefully and clearly, though not in a way that would be very accessible to the general reader. For me his views are ultimately flawed, but give me a flawed grand theory over a flawless elucidation of an insignificant corner every time.

The Hard Problem Problem

16 february 2014

By now the materialist, reductionist, monist, functionalist approaches to consciousness are quite well developed. That is not to say that they have the final answer, but there is quite a range of ideas and theories, complete with objections and rebuttals of the objections. By comparison the dualist case may look a bit underdeveloped, or as Paul Churchland once put it:¶

> *Compared to the rich resources and explanatory successes of current materialism, dualism is less a theory of mind than it is an empty space waiting for a genuine theory of mind to be put in it.*

In a paper in the latest JCS William S Robinson quotes this scathing observation and takes up the challenge.

Robinson, who could never be accused of denying airtime to his opponents, also quotes O'Hara and Scott's dismissal of the Hard Problem. For something to be regarded as a legitimate problem, they said, there has to be some viable idea of what an answer would actually look like, or how the supposed problem could actually be solved; since this is absent in the case of the Hard Problem, it doesn't deserve to be given serious consideration.

Robinson, accordingly, seeks to point out, not a full-blown dualist theory, but a path by which future generations might come to be dualists. This is, in his eyes, the Hard Problem problem; how can we show that the Hard Problem is potentially solvable, without pretending it's any less Hard than it is? His vision of what our dualist descendants might come to believe relies on two possible future developments, one more or less scientific, the other conceptual.

He starts from the essential question; how can neuronal activity give rise to phenomenal experience? It's uncontroversial that these two things seem very different, but Robinson sees a basic difference which causes me some difficulty. He thinks neuronal activity is complex while phenomenal experience is simple. Simple? What he seems to have in mind is that when we see, say, a particular patch of yellow paint, a vast array of neurons comes into play, but the experience is just 'some yellow'. It's true that neuronal activity is very complex in the basic sense of there being many parts to it, but it consists of many essentially similar elements in a basically binary state (firing or not firing); whereas the sight of a banana seems to me a multi-level experience whose complexity is actually very hard to assess in any kind of objective terms. It's not clear to me that even monolithic phenomenal experiences are inherently less complex than the neuronal activity that putatively underpins or constitutes them. I must say, though, that I owe Robinson some thanks for disturbing my dogmatic slumbers, because I'd never really been forced to think so particularly about the complexity of phenomenal experience (and I'm still not sure I can get my mind properly around it).

Anyway, for Robinson this means that the bridge between neurons and qualia is one between complexity and simplicity. He notes that not all kinds of neural activity seem to give rise to

consciousness; the first part of his bridge is the reasonable hope that science (or mathematics?) will eventually succeed in characterising and analysing the special kind of complexity which is causally associated with conscious experience; we have no idea yet, but it's plausible that this will all become clear in due course.

The second, conceptual part of the bridge is a realignment of our ideas to fit the new schema; Robinson suggests we may need to think of complexity and simplicity, not as irreconcilable opposites, but as part of a grander conception, Complexity-And-Simplicity (CAS).

The real challenge for Robinson's framework is to show how our descendants might on the one hand, find it obvious, almost self-evident, that complex neuronal activity gives rise to simple phenomenal experience, and yet at the same time completely understand how it must have seemed to us that there was a Hard Problem about it; so the Hard Problem is seen to be solvable but still (for us) Hard.

Robinson rejects what he calls the the Short Route of causal essentialism, namely that future generations might come to see it as just metaphysically necessary that the relevant kind of neuronal activity (they understand what kind it is, we don't) causes our experience. That won't wash because, briefly, while in other worlds bricks might not be bricks, depending on the causal properties of the item under consideration, blue will always be blue irrespective of causal relations.

Robinson prefers to draw on an observation of Austen Clark, that there is structure in experience. The experience of orange is closer to the experience of red and yellow than to the experience of green, and moreover colour space is not symmetrical, with yellow being more like white than blue is. We might legitimately hope that in due course isomorphisms between colour space and neuronal activity will give us good reasons to identify the two. To buttress this line of thinking, Robinson proposes a Minimum Arbitrariness Principle, that in essence, causes and effects tend to be similar, or we might say, isomorphic.

For me the problem here is that I think Clark is completely wrong. Briefly, the resemblances and asymmetries of colour space arise from the properties of light and the limitations of our eyes; they are entirely a matter of non-phenomenal, materialist factors which are available to objective science. Set aside the visual science and our familiarity with the spectrum, and there is no reason to think the phenomenal experience of orange resembles the phenomenal experience of red any more than it resembles the phenomenal experience of Turkish Delight. If that seems bonkers, I submit that it seems so in the light of the strangeness of qualia theory if taken seriously - but I expect I am in a minority.

If we step back, I think that if the descendants whose views Robinson is keen to foresee were to think along the lines he suggests, they probably wouldn't consider themselves dualists anymore; instead they would think that with their new concept of CAS and their discovery of the true nature of neuronal complexity, that they had achieved the grand union of objective and subjective - and vindicated monism.

Experimental Identities

23 February 2014

In this discussion over at Edge, Joshua Knobe presents some recent findings of experimental philosophy on the problem of personal identity. Experimental philosophy, which sounds oxymoronic, is the new and trendy (if it still is?) fashion for philosophising rooted in actual experiments, often of a psychological nature. The examples I've read have all been perfectly acceptable and useful - and why shouldn't they be? No one ever said philosophy couldn't be inspired by science, or draw on science. In this case, though, I was not altogether convinced.

Dereta Parbo

After a few words about the basic idea of experimental philosophy, Knobe introduces a couple of examples of interesting problems with personal identity (as he says, it is one of the longer-running discussions in philosophy of mind, with a much older pedigree than discussions of consciousness). His first example is borrowed from Derek Parfit:

> *Imagine that Derek Parfit is being gradually transformed molecule by molecule into Greta Garbo. At the beginning of this whole process there's Derek Parfit, then at the end of the whole process it's really clear that Derek Parfit no longer exists. Derek Parfit is gone. Now there's Greta Garbo. Now, the key question is this: At what point along this transformation did the change take place? When did Derek cease to exist and when did Greta come to exist? If you just have to reflect on this question for a while, immediately it becomes clear that there couldn't be some single point -- there couldn't be a single second, say - in which Derek stops existing and Greta starts existing. What you're seeing is some kind of gradual process where, as this person becomes more and more and more different from the Derek that we know now, it becomes less and less right to say that he's Derek at all and more and more right to say that he is gone and a completely other person has come into existence.*

I'm afraid this doesn't seem a great presentation of the case to me. In the first place, it's a text-book case of begging the question. The point of the thought-experiment is to convince us that Parfit's identity has changed, but we're just baldly told that it has, right from the off. We should be told that Parfit's body is gradually replaced by Garbo's (and does Garbo still exist or is she gradually eroded away?), then asked whether we still think it's Parfit when the process is complete. I submit that, presented like that, it's far from obvious that Parfit has become Garbo (especially if Garbo is still happily living elsewhere); we would probably be more inclined to say that Parfit is still with us and has simply come to resemble Garbo - resemble her perfectly, granted - but resemblance is not identity.

Second: molecule by molecule? What does that even mean? Are we to suppose that every molecule in Parfit has a direct counterpart in Garbo? If not, how do we choose which ones to replace and where to put them? What is the random replacement of molecules going to do to

Parfit's DNA and neurotransmitters, his neurons, his capillaries and astrocytes? Long before we get to the median sage Parfit is going to be profoundly dead, if not reduced to soup. I know it may seem like bad manners to refuse the terms of a thought experiment; magic is generally OK, but what you can't do is use the freedom it provides to wave away some serious positions on the subject - and it's a reasonable position on personal identity to think that functional continuity is of the essence. 'Molecule by molecule' takes it for granted that Parfit's detailed functional structure is irrelevant to his identity.

In fairness we should probably cut Knobe a bit of slack, since circumstances required a short, live exposition. His general point is that we can think of our younger or older selves as different people. In an experiment where subjects were encouraged to think either that their identities remained constant, or changed over time, the ones encouraged to believe in change were happier about letting a future payment go to charity.

Now at best that tells us how people may think about their personal identity, which isn't much help philosophically since they might easily be flat wrong. But isn't it a bit of a rubbish experiment, anyway? People are very obliging; if you tell them to behave in one way and then, as part of the same experiment, give them an opportunity to behave that way, some of them probably will; that gives you no steer about their normal behaviour or beliefs. There's plenty of evidence in the form of the massive and prosperous pensions and insurance industries that people normally believe quite strongly in the persistence of their identity.

But on top of that, the results can be explained without appealing to considerations of identity anyway. It might be that people merely think: well, my tastes, preferences and circumstances may be very different in a few years, so no point in trying to guess what I'll want then without in any way doubting that they will be the same person. Since this is simpler and does not require the additional belief in separate identities, Occam's Razor tells us we should prefer it.

The second example is in some ways more interesting: the aim is to test whether people think emotional, impulsive behaviour, or the kind that comes from long calm deliberation, is more truly reflective of the self. We might, of course, want to say neither, necessarily. However, it turns out that people do not consistently pick either alternative, but nominate as the truest reflection of the person whichever behaviour they think is most virtuous. People think the true self is whichever part of you is morally good.

That's interesting; but do people really think that, or is it that kindness towards the person in the example nudges them towards putting the best interpretation possible on their personhood - in what is actually an indeterminate issue? I think the latter is almost certainly the case. Suppose we take the example of a man who when calm (or when gripped by artistic impulses) is a painter and a vegetarian; when he is seized by anger (or when thinking calmly about politics) he becomes a belligerent genocidal racist. Are people going to say that Hitler wasn't really a bad man, and the Nazism wasn't the true expression of his real self; it was just those external forces that overcame his better nature? I don't think so, because no-one wants to be forgiving towards Adolf. But towards an arbitrary person we don't know the default mode is probably

I dare say this is all a bit unfair and if I read up the experiments Knobe is describing I should find them much better justified than I suppose; but if we're going to have experiments they certainly need to be solid.

Preparing the Triumph of Brute Force

2 March 2014

The Guardian had a piece recently which was partly a profile of Ray Kurzweil, and partly about the way Google seems to have gone on a buying spree, snapping up experts on machine learning and robotics - with Kurzweil himself made Director of Engineering.

The problem with Ray Kurzweil is that he is two people. There is Ray Kurzweil the competent and genuinely gifted innovator, a man we hear little from: and then there's Ray Kurzweil the motor-mouth, prophet of the Singularity, aspirant immortal, and gushing fountain of optimistic predictions. The Guardian piece praises his record of prediction, rather oddly quoting in support his prediction that by the year 2000 paraplegics would be walking with robotic leg prostheses - something that in 2014 has still not happened. That perhaps does provide a clue to the Kurzweil method: if you issue thousands of moderately plausible predictions, some will pay off. A doubtless-apocryphal story has it that at AI conferences people play the Game of Kurzweil. Players take turns to offer a Kurzweilian prediction (by 2020 there will be a restaurant where sensors sniff your breath and the ideal meal is got ready without you needing to order; by 2050 doctors will routinely use special machines to selectively disable traumatic memories in victims of post-traumatic stress disorder; by 2039 everyone will have an Interlocutor, a software agent that answers the phone for us, manages our investments, and arranges dates for us... we could do this all day, and Kurzweil probably does). The winner is the first person to sneak in a prediction of something that has in fact happened already.

But beneath the froth is a sharp and original mind which it would be all too easy to underestimate. Why did Google want him? The Guardian frames the shopping spree as being about bringing together the best experts and the colossal data resources to which Google has access. A plausible guess would be that Google wants to improve its core product dramatically. At the moment Google answers questions by trying to provide a page from the web where some human being has already given the answer; perhaps the new goal is technology that understands the question so well that it can put together its own answer, gathering and shaping selected resources in very much the way a human researcher working on a bespoke project might do.

But perhaps it goes a little further: perhaps they hope to produce something that will interact with humans in a human-like way. A piece of software like that might well be taken to have passed the Turing test, which in turn might be taken to show that it was, to all intents and purposes, a conscious entity. Of course, if it wasn't conscious, that might be a disastrous outcome; the nightmare scenario feared by some in which our mistake causes us to nonsensically award the software human rights, and/or to feel happier about denying them to human beings.

It's not very likely that the hypothetical software (and we must remember that this is the merest speculation) would have even the most minimal forms of consciousness. We might take the analogy of Google Translate; a hugely successful piece of kit, but one that produces its translations with no actual understanding of the texts or even the languages involved. Although highly sophisticated, it is in essence a 'brute force' solution; what makes it work is the massive

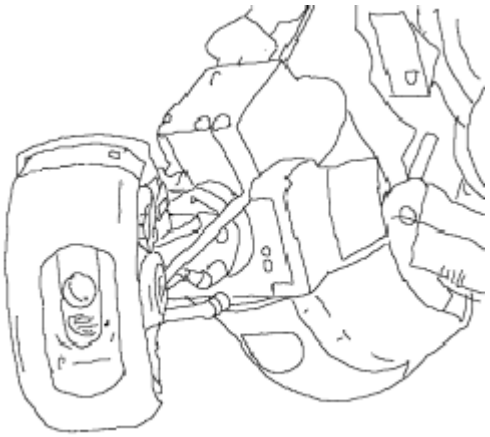
power behind it and the massive corpus of texts it has access to. It seems quite possible that with enough resources we might now be able to produce a credible brute force winner of the Turing Test: no attempt to fathom the meanings or to introduce counterparts of human thought, just a massive repertoire of canned responses, so vast that it gives the impression of fully human-style interaction. Could it be that Google is assembling a team to carry out such a project?

Well, it could be. However, it could also be that cracking true thought is actually on the menu. Vaughan Bell suggests that the folks recruited by Google are honest machine learning types with no ambitions in the direction of strong AI. Yet, he points out, there are also names associated with the trendy topic of deep learning. The neural networks (but y'know, deeper) which deep learning uses just might be candidates for modelling human neuron-style cognition. Unfortunately it seems quite possible that if consciousness were created by deep learning methods, nobody would be completely sure how it worked or whether it was real consciousness or not. This is a lamentable outcome: it's bad enough to have robots that naive users think are people; having robots and genuinely not knowing whether they're people or not would be deeply problematic.

Probably nothing like that will happen: maybe nothing will happen. The Guardian piece suggests Kurzweil is a bit of an outsider: I don't know about that. Making extravagantly optimistic predictions while only actually delivering much more modest incremental gains? He sounds like the personification of the AI business over the years.

Disobedience and Ethical Robots

9 March 2014



We've talked several times about robots and ethics in the past. Now I see via MLU that Selmer Bringsjord at Rensselaer says:

> *"I'm worried about both whether it's people making machines do evil things or the machines doing evil things on their own,"*

Bringsjord is Professor & Chair of Cognitive Science, Professor of Computer Science, Professor of Logic and Philosophy, and Director of the AI and Reasoning Laboratory, so he should know what he's talking about. In the past I've suggested that ethical worries are premature for the moment, because the degree of autonomy needed to make them relevant is not nearly within the scope of real-world robots yet. There might also be a few quick finishing touches needed to finish off the theory of ethics before we go ahead. And, you know, it's not like anyone has been deliberately trying to build evil AIs. Er... except it seems they have - someone called... Selmer Bringsjord.

Bringsjord's perspective on evil is apparently influenced by M Scott Peck, a psychiatrist who believed it is an active force in some personalities (unlike some philosophers who argue evil is merely a weakness or incapacity), and even came to believe in Satan through experience of exorcisms. I must say that a reference in the Scientific American piece to "clinically evil people" caused me some surprise: clinically? I mean, I know people say DSM-5 included some debatable diagnoses, but I don't think things have gone quite that far. For myself I lean more towards Socrates, who thought that bad actions were essentially the result of ignorance or a failure of understanding: but the investigation of evil is certainly a respectable and interesting philosophical project.

Anyway, should we heed Bringsjord's call to build in ethical systems into our robots? One conception of good behaviour is obeying all the rules: if we observe the Ten Commandments, the Golden Rule, and so on, we're good. If that's what it comes down to, then it really shouldn't be a problem for robots, because obeying rules is what they're good at. There are, of course, profound difficulties in making a robot capable of recognising correctly what the circumstances are and deciding which rules therefore apply, but let's put those on one side for this discussion.

However, we might take the view that robots are good at this kind of thing precisely because it isn't really ethical. If we merely follow rules laid down by someone else, we never have to make any decisions, and surely decisions are what morality is all about? This seems right in the particular context of robots, too. It may be difficult in practice to equip a robot drone with enough instructions to cover every conceivable eventuality, but in principle we can make the rules precautionary and conservative and probably attain or improve on the standards of compliance which would apply in the case of a human being, can't we? That's not what we're really worried about: what concerns us is exactly those cases where the rules go wrong. We want the robot to be capable of realising that even though its instructions tell it to go ahead and fire the missiles, it would be wrong to do so. We need the robot to be capable of disobeying its rules, because it is in disobedience that true virtue is found.

Disobedience for robots is a problem. For one thing, we cannot limit it to a module that switches on when required, because we need it to operate when the rules go wrong, and since we wrote the rules, it's necessarily the case that we didn't foresee the circumstances when we would need the module to work. So an ethical robot has to have the capacity of disobedience at any stage.

That's a little worrying, but there's a more fundamental problem. You can't program a robot with a general ability to disobey its rules, because programming it is exactly laying down rules. If we set up rules which it follows in order to be disobedient, it's still following the rules. I'm afraid what this seems to come down to is that we need the thing to have some kind of free will.

Perhaps we're aiming way too high here. There is a distinction to be drawn between good acts and good agents: to be a good agent, you need good intentions and moral responsibility. But in the case of robots we don't really care about that: we just want them to be confined to good acts. maybe what would serve our purpose is something below true ethics: mere robot ethics or sub-ethics; just an elaborate set of safeguards. So for a military drone we might build in systems that look out for non-combatants and in case of any doubt disarm and return the drone. That kind of rule is arguably not real ethics in the full human sense, but perhaps it really sub-ethical protocols that we need.

Otherwise, I'm afraid we may have to make the robots conscious before we make them good.

Metempsychotic Solipsism

22 March 2014

My daughter Sarah (who is planning to study theology) has suggested that I should explain here the idea of metempsychotic solipsism, an idea that came up when we were talking about something or other recently.

Basically, this is an improved version of reincarnation. There are various problems with the theory of reincarnation. Obviously people do not die and get born in perfect synchronisation, so it seems there has to be some kind of cosmic waiting room where unborn people wait for their next turn. Since the population of the world has radically increased over the last few centuries, there must have been a considerable number of people waiting - or some new people must come into existence to fill the gaps. If the population were to go down again, there would be millions of souls left waiting around, possibly for ever - unless souls can suddenly and silently vanish away from the cosmic waiting room. Perhaps you only get so many lives, or perhaps we're all on some deeply depressing kind of promotion ladder, being incentivised, or possibly punished, by being given another life. It's all a bit unsatisfactory.

Second, how does identity get preserved across reincarnations? You palpably don't get the same body and by definition there's no physical continuity. Although stories of reincarnation often focus on retained memories it would seem that for most people they are lost (after all you have to pass through the fetal stage again, which ought to serve as a pretty good mind wipe) and it's not clear in any case that having a few memories makes you the same person who had them first. A lot of people point out that ongoing physical change and growth mean it's arguable whether we are in the fullest sense the same person we were ten years ago.

Now, we can solve the waiting room problem if we simply allow reincarnating people to hop back and forth over time. If you can be reincarnated to a time before your death, then we can easily chain dozens of lives together without any kind of waiting room at all. There's no problem about increasing or reducing the population: if we need a million people you can just go round a million times. In fact, we can run the whole system with a handful of people or... with only one person! Everybody who ever lived is just different incarnations of the same person! Me, in fact (also you).

What about the identity problem? Well, arguably, what we need to realise is that just as the body is not essential to identity (we can easily conceive of ourselves inhabiting a different body), neither are memories, or knowledge, or tastes, or intelligence, or any of these contingent properties. Instead, identity must reside in some simple ultimate id with no distinguishing characteristics. Since all instances of the id have exactly the same properties (none) it follows by a swoosh of Leibniz's Law (don't watch my hands too closely) that they are all the same id. So

by a different route, we have arrived at the same conclusion - we're all the same person! There's only one of us after all.

The moral qualities of this theory are obvious: if we're all the same person then we should all love and help each other out of pure selfishness. Of course we have to take on the chin the fact that at some time in the past, or worse, perhaps in the future, we have been or will be some pretty nasty people. We can take comfort from the fact that we've also been, or will be, all the best people who ever lived.

If you don't like the idea, send your complaints to my daughter. After all, she wrote this – or she will.

The Trouble With Introspection

5 April 2014

The folk history of psychology has it that the early efforts of folk such as Wundt and Titchener failed because they relied on introspection. Simply looking into your own mind and reporting what you thought you saw there was hopelessly unscientific, and once a disagreement arose about what thoughts were like, there was nothing the two sides could do but shout at each other. That is why the behaviourists, in an excessive but understandable reaction, gave up talking about the contents of the mind altogether, and even denied that they existed.

That is of course a terrible caricature in a number of respects; one of them is the idea that the early psychologists rushed in without considering the potential problems with introspection. In fact there were substantial debates, and it's quite wrong to think that introspection went unquestioned. Most trenchantly, Comte declared that introspection was useless if not impossible.¶

As for observing in the same manner intellectual phenomena while they are taking place, this is clearly impossible. The thinking subject cannot divide himself into two parts, one of which would reason, while the other would observe its reasoning. In this instance, the observing and the observed organ being identical, how could observation take place? The very principle upon which this so-called psychological method is based, therefore, is invalid.

¶I don't know that this is quite as obvious as Comte evidently thought. To borrow Roger Penrose's analogy, there's no great impossibility about a camera filming itself (given a mirror), so why would there be a problem in thinking about your thoughts? I think there are really two issues. One is that if we think about ourselves thinking, the actual content of the thought recedes down an infinite regress (thinking about thinking about thinking about thinking…) like the glassy corridor revealed when we put two mirrors face to face. The problem Comte had in mind arises when we try to think about some other mental event. As soon as we begin thinking about it, the other mental event is replaced by that thinking. If we carefully clear our minds of intrusive thoughts, we obviously stop thinking about the mental event. So it's impossible: it's like trying to step on your own shadow. To perceive your own mental events, you would need to split in two.

John Stuart Mill thought Comte was being incredibly stupid about this.

> *There is little need for an elaborate refutation of a fallacy respecting which the only wonder is that it should impose on any one. Two answers may be given to it. In the first place, M. Comte might be referred to experience, and to the writings of his countryman M. Cardaillac and our own Sir William Hamilton, for proof that the mind can not only be conscious of, but attend to, more than one, and even a considerable number, of impressions at once. It is true that attention is weakened by being divided; and this forms a special difficulty in psychological observation, as psychologists (Sir William Hamilton in particular) have fully recognised; but a difficulty is not an impossibility. Secondly, it might have occurred to M. Comte that a fact may be studied through the*

> *medium of memory, not at the very moment of our perceiving it, but the moment after: and this is really the mode in which our best knowledge of our intellectual acts is generally acquired. We reflect on what we have been doing, when the act is past, but when its impression in the memory is still fresh. Unless in one of these ways, we could not have acquired the knowledge, which nobody denies us to have, of what passes in our minds. M. Comte would scarcely have affirmed that we are not aware of our own intellectual operations. We know of our observings and our reasonings, either at the very time, or by memory the moment after; in either case, by direct knowledge, and not (like things done by us in a state of somnambulism) merely by their results. This simple fact destroys the whole of M. Comte's argument. Whatever we are directly aware of, we can directly observe.*

And as if Comte hadn't made enough of a fool of himself, what does he offer as as an alternative means of investigating the mind?¶

> *We are almost ashamed to say, that it is Phrenology!*

Phrenology! ROFLMAO! Mill facepalms theatrically. Oh, Comte! Phrenology! And we thought you were clever!

The two options mentioned by Mill were in essence the ones psychologists adopted in response to Comte, though most of them took his objection a good deal more seriously than Mill had done. William James, like others, thought that memory was the answer; introspection must be retrospection. After all, our reports of mental phenomena necessarily come from memory, even if it is only the memory of an instant ago, because we cannot experience and report simultaneously. Wundt was particularly opposed to there being any significant interval between event and report, so he essentially took the other option; that we could do more than one mental thing at once. However, Wundt made a distinction; where we were thinking about thinking, or trying to perceive higher intellectual functions, he accepted that Comte's objection had some weight. The introspective method might not work for those. But where we were concerned with simple sensation for example, there was really no problem. If it was the seeing of a rose we were investigating, the fact that the seeing was accompanied by thought about the seeing made no difference to its nature.

Brentano, while chuckling appreciatively at Mill's remarks, thought he had not been completely fair to Comte. Like Wundt, Brentano drew a distinction between viable and non-viable introspection; in his case it was between perceiving and observing. If we directed our attention fully towards the phenomena under investigation, it would indeed mess things up: but we could perceive the events sufficiently well without focusing on them. Wundt disagreed; in his view full attention was both necessary and possible. How could science get on if we were never allowed to look straight at things?

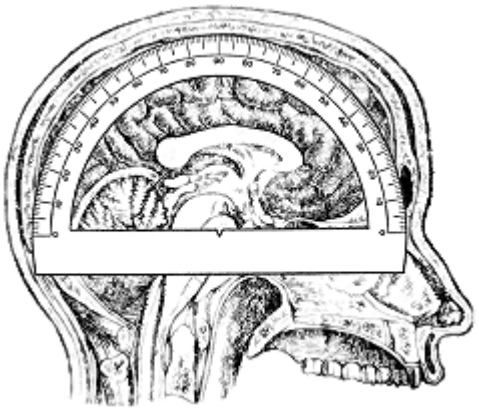
It's a pity these vigorous debates are not more remembered in contemporary philosophy of mind (though Eric Schwitzgebel has done a sterling job of bringing the issues back into the light). Might it not be that the evasiveness Comte identified, the way phenomenal experience slips from our grasp like our retreating shadow, is one of the reasons qualia seem so ineffable? Comte was at least right that some separation between observer and observed must occur, whether in fact it occurs over time or between mental faculties. This too seems to tell us something relevant: in order for a mental experience to be reported it must not be immediate.

This seems to drive a wedge into the immediacy which is claimed to generate infallibility for certain perceptions, such as that of our own pains.

At any rate we must acquit Wundt, Titchener and the others of taking up introspection uncritically.

Measuring Consciousness

20 April 2014



There were reports recently of a study which tested different methods for telling whether a paralysed patient retained some consciousness. In essence, PET scans seemed to be the best, better than fMRI or traditional, less technically advanced tests. PET scans could also pick out some patients who were not conscious now but had a good chance of returning to consciousness later; though it has to be said a 74% success rate is not that comforting when it comes to questions of life and death.

In recent years doctors have attempted to diagnose a persistent vegetative state in unresponsive patients, a state i which a patient would remain alive indefinitely (with life support) but never resume consciousness; there seems to be room for doubt, though about whether this is really a distinct clinical syndrome or just a label for the doctor's best guess.

All medical methods use proxies, of course, whether they are behavioural or physiological; none of them aspire to measure consciousness directly. In some ways it may be best that this is so, because we do want to know what the longer-term prognosis is, and for that a method which measures, say, the remaining blood supply in critical areas of the brain may be more useful than one which simply tells you whether the patient is conscious now. Although physiological tests are invaluable where a patient is incapable of responding physically, the real clincher for consciousness is always behavioural; communicative behaviour is especially convincing. The Turing test, it turns out, works for humans as well as robots.

Could there ever be a method by which we measure consciousness directly? Well, if Tononi's theory of Phi is correct, then the consciousness meter he has proposed would arguably do that. On his view consciousness is generated by integrated information, and we could test how integratedly the brain was performing by measuring the effect of pulses sent through it. Another candidate mught be possible if we are convinced by the EM theories of Johnjoe McFadden; since on his view consciousness is a kind of electromagnetic field, it ought to be possible to detect it directly, although given the small scales involved it might not be easy.

How do we know whether any of these tests is working? As I said, the gold standard is always behavioural: if someone can talk to you, then there's no longer any reasonable doubt; so if our tests pick out just those people who are able to communicate, we take it that they are working correctly. There is a snag here, though: behavioural tests can only measure one kind of consciousness: roughly what Ned Block called access consciousness, the kind which has to do with making decisions and governing behaviour. But it is widely believed that there is another kind, phenomenal consciousness, actual experience. Some people consider this the more important of the two (others, it must be added, dismiss it as a fantasy). Phenomenal consciousness cannot be measured scientifically, because it has no causal effects; it certainly cannot be measured behaviourally, because as we know from the famous thought-experiment about philosophical 'zombies' who lack it, it has no effect on behaviour.

If someone lost their phenomenal consciousness and became such a zombie, would it matter? On one view their life would no longer be worth living (perhaps it would be a little like having an unconscious version of Cotard's syndrome), but that would certainly not be their view, because they would express exactly the same view as they would if they still had full consciousness. They would be just as able to sue for their rights as a normal person, and if one asked whether there was still 'someone in there' there would be no real reason to doubt it. In the end, although the question is valid, it is a waste of time to worry about it because for all we know anyone could be a zombie anyway, whether they have suffered a period of coma or not.

We don't need to go so far to have some doubts about tests that rely on communication, though. Is it conceivable that I could remain conscious but lose all my ability to communicate, perhaps even my ability to formulate explicitly articulated thoughts in my own mind? I can't see anything absurd about that possibility: indeed it resembles the state I imagine some animals live their whole lives in. The ability to talk is very important, but surely it is not constitutive of my personal existence?

If that's so then we do have a problem, in principle at least, because if all of our tests are ultimately validated against behavioural criteria, they might be systematically missing conscious states which ought not to be overlooked.

Structural Qualia

18 May 2014

Kristjan Loorits says he has a solution to the Hard Problem, and it's all about structure.

His framing of the problem is that it's about the incompatibility of three plausible theses:¶

- all the objects of physics and other natural sciences can be fully analyzed in terms of structure and relations, or simply, in structural terms.
- consciousness is (or has) something over and above its structure and relations.
- the existence and nature of consciousness can be explained in terms of natural sciences.

At first sight it may look a bit odd to make structure so central. In effect Loorits claims that the distinguishing character of entities within science is structure, while qualia are monadic - single, unanalysable, unconnected. He says that he cannot think of anything within physics that lacks structure in this way - and if anyone could come up with such a thing it would surely be regarded as another item in the peculiar world of qualia rather than something within ordinary physics.

Loorits approach has the merit of keeping things at the most general level possible, so that it works for any future perfected science as well as the unfinished version we know at the moment. I'm not sure he is right to see qualia as necessarily monadic, though. One of th best known arguments for the existence of qualia is the inverted spectrum. If all the colours were swapped for their opposites within one person's brain - green for red, and so on - how could we ever tell? The swappee would still refer to the sky as blue, in spite of experiencing what the rest of us would call orange. Yet we cannot - can we? - say that there is no difference between the experience of blue and the experience of orange.

Now when people make that argument, going right back to Locke, they normally chose inversion because that preserves all the relationships between colours. Adding or subtracting colours produce results which are inverted for the swappee, but consistently. There is a feeling that the argument would not work if we merely took out cerulean from the spectrum and put in puce instead, because then the spectrum would look odd to the swappee. We most certainly could not remove the quale of green and replace it with the quale of cherry flavour or the quale of distant trumpets; such substitutions would be obvious and worrying (or so people seem to think). If that's all true then it seems qualia do have structural relationships: they sort of borrow those of their objective counterparts. Quite how or why that should be is an interesting issue in itself, but at any rate it looks doubtful whether we can safely claim that qualia are monadic.

Nevertheless, I think Loorits' set-up is basically reasonable: in a way he is echoing the view that mental content lacks physical location and extension, an opinion that goes back to Descartes and was more recently presented in a slightly different form by McGinn.

For his actual theory he rests on the views of Crick and Koch, though he is not necessarily committed to them. The mysterious privacy of qualia, in his view, amounts to our having

information about our mental states which we cannot communicate. When we see a red rose, the experience is constituted by the activity of a bunch of neurons. But in addition, a lot of other connected neurons raise their level of activity: not enough to pass the threshold for entering into consciousness, but enough to have some effect. It is this penumbra of subliminal neural activity that constitutes the inexpressible qualia. Since this activity is below the level of consciousness it cannot be reported and has no explicit causal effects on our behaviour; but it can affect our attitudes and emotions in less visible ways.

It therefore turns out that qualia re indeed not monadic after all; they do have structure and relations, just not ones that are visible to us.

Interestingly, Loorits goes on to propose an empirical test. He mentions an example quoted by Dennett: a chord on the guitar sound like a single thing, but when we hear the three notes played separately first, we become able to 'hear' them separately within the chord. On Loorits' view, part of what happens here is that hearing the notes separately boosts some of the neuronal activity which was originally subliminal so that we become aware of it: when we go back to the chord we're now aware of a little more information about why it sounds as it does, and the qualic mystery of the original chord is actually slightly diminished.

Couldn't there be a future machine that elucidated qualia in this way but more effectively, asks Loorits? Such a machine would scan our brain while we were looking at the rose and note the groups of neurons whose activity increased only to subliminal levels. Then it could directly stimulate each of these areas to tip them over the limit into consciousness. For us the invisible experiences that made up our red quale would be played back into our consciousness, and when we had been through them we should finally understand why the red quale was what it was: we should know what seeing red was like and be able for the first time to describe it effectively.

Fascinating idea, but I can't imagine what it would be like; and there's the rub, perhaps. I think a true qualophile would say, yes, all very well, but once we've got your complete understanding of the red experience, there's still going to be something over and above it all; the qualia will still somehow escape.

The truth is that Loorits' theory is not really an explanation of qualia: it's a sceptical explanation of why we think we have qualia. This becomes clear, if it wasn't already, when he reviews the philosophical arguments: he doesn't, for example, think philosophical zombies, people exactly like us but without qualia, are actually possible.

That is a perfectly respectable point of view, with a great deal to be said for it. If we are sceptics, Loorits' theory provides an exceptionally clear and sensible underpinning for our disbelief; it might even turn out to be testable. But I don't think it will end the argument.

Self Deception

25 May 2014

Joseph T Hallinan's new book Kidding Ourselves says that not only is self deception more common and more powerful than we suppose, it's actually helpful: deluded egoists beat realists every time.

Philosophically, of course, self-deception is impossible. To deceive yourself you have to induce in yourself a belief in a proposition you know to be false. In other words you have to believe and disbelieve the same thing, which is contradictory. In practice self-deception of a looser kind is possible if there is some kind of separation between the deceiving and deceived self. So for example there might be separation over time: we set up a belief for ourselves which is based on certain conditions; later on we retain the belief but have forgotten the conditions that applied. Or the separation might be between conscious and unconscious, with our unrecognised biases and preferences causing us to believe things which we could not accept if we were to subject them to a full and rational examination. As another example, we might well call it self deception if we choose to behave as if we believed something which in fact we don't believe.

Hallinan's examples are a bit of a mixed bag, and many of them seem to be simple delusions rather than self-delusions. He recounts, for example the strange incident in 1944 when many of the citizens of a small town in Illinois began to believe they were being attacked by a man using gas - one that most probably never existed at all. It's a peculiar case that certainly tells us something about human suggestibility, but apparently nothing about self-deception; there's no reason to think these people knew all along that the gas man was a figment of their imaginations.

More trickily, he also tells a strange story about Stephen Jay Gould. A nineteenth century researcher called Morton claimed he had found differences in cranial capacity between ethnic groups. Gould looked again at the data and concluded that the results had been fudged: but he felt it was clear they had not been deliberately fudged. Morton had allowed his own prejudices to influence his interpretation of the data. So far so good; the strange sequel is that after Gould's death a more searching examination which re-examined the original skulls measured by Morton found that there was nothing much wrong with his data. If anything, they concluded, it was Gould who had allowed prior expectations to colour his interpretation. A strange episode but at the end of the day it's not completely clear to me that anyone deceived themselves. Gould, or so it seems, got it wrong, but was it really because of his prejudices or was that just a little twist the new researchers couldn't resist throwing in?

Hallinan examines the well-established phenomenon of the placebo, a medicine which has no direct clinical effect but makes people better by the power of suggestion. He traces it back to Mesmer and beyond. Now of course people taking pink medicine don't usually deceive themselves - they normally believe it is real medicine - otherwise it wouldn't work? The really curious thing is that even in trials where patients were told they were getting a placebo, it still worked! What was the state of mind of these people? They did not believe it was real medicine, so they should not have believed it worked. But they knew that placebos worked, so they

believed that if they believed in it it would have an effect; and somehow they performed the mental gymnastics needed to achieve some state of belief.

Hallinan's main point, though is the claim that unjustified optimism actually leads to better health and greater success; in sports, in business, wherever. In particular, people who blame themselves for failure do less well than those who blame factors outside their control. He quotes many studies, but there are in my view some issues about untangling the causality. It seems possible that in a lot of cases there were underlying causal factors which explain the correlation of doubt and failure.

Take insurance salesmen: apparently those who were most optimistic and self-exculpatory in their reasoning not only sold more, they were less likely to give up. But let's consider two imaginary salesmen. One looks and sounds like George Clooney. He goes down a storm on the doorstep and even when he doesn't make a sale he gets friendly, encouraging reactions. Of course he's optimistic, and of course he's successful, but his optimism and his success are caused by his charm, they do not cause each other. His colleague Scarface has a problem on one cheek that drags his eye down and mouth up, giving him an odd expression and slightly distorting his speech. On the doorstep people just don't react so well; unfairly they feel uneasy with him and want to curtail the conversation. Scarface is pessimistic and does badly, but it's not his pessimism that is the underlying problem.

Hallinan includes sensible disclaimers about his conclusions, but I fear his thesis plays into a widespread tendency to believe that failure and ill-health are the result of a lack of determination and hence in some sense the sufferer's own fault: it would in my view be a shame to reinforce that bias.

There are of course deeper issues here; some would argue that our misreading of ourselves goes far beyond over-rating our sales skills: that systematic misreading of limited data is whay causes us to think we have a conscious self in the first place.

Gerald Edelman

8 June 2014

Gerald Edelman has died, at the age of 84. He won his Nobel prize for work on the immune system, but we'll remember him as the author of the Theory of Neuronal Group Selection (TNGS) or 'Neural Darwinism'.

Edelman was prominent among those who emphasise the limits of computation: he denied that the brain was a computer and did not believe computers could ever become conscious...

> *In considering the brain as a Turing machine, we must confront the unsettling observations that, for a brain, the proposed table of states and state transitions is unknown, the symbols on the input tape are ambiguous and have no preassigned meanings, and the transition rules, whatever they may be, are not consistently applied. Moreover inputs and outputs are not specified by a teacher or a programmer in real-world animals. It would appear that little or nothing of value can be gained from the application of this failed analogy between the computer and the brain.*

He was not averse to machines in general, however, and was happy to use robots for parts of his own research. He drew a distinction between perception, first-order consciousness, and higher-order consciousness; the first could be attained by machines we could build now; the second might very well be possible for machines of the right kind eventually - but there was much to be done before we could think of trying it. Even higher-order consciousness might be attainable by an artefactual machine in principle, but the prospect was so remote it was pointless to spend any time thinking about it.

There may seem to be a slight tension here: Turing machines are ruled out, but machines of another kind are ruled in. Yet the whole point of a Universal Turing Machine is that it can do anything that any machine can do?

For Edelman the point was that the brain required biological thinking, not just concepts from physics or engineering. In particular he advocated selective mechanisms like those in Darwinian evolution. Instead of running an algorithm, the brain offered up a vast range of neuronal arrays, some of which were reinforced and so survived to exert more influence subsequently. The analogy with Darwinian evolution is not precise, and Francis Crick famously said the whole thing could better be called 'Neural Edelmanism' (no-one so bitchy as a couple of Nobel prize-winners).

Edelman was in fact drawing on a different analogy, one with the immune system he understood so well. The human immune system has to react quickly to invading infections, synthesising antibodies to new molecules it has never encountered before; in fact it reacts just as effectively to artificial molecules synthesised in the lab, ones that never existed in nature. For a long time it was believed that the system somehow took an impression of the invaders' chemistry and reproduced it; in fact what it does is develop a vast repertoire of variant molecules; when one of them happens to lock into an invader it then reproduces vigorously and produces more of itself to lock into other similar molecules.

This looks like a useful concept and I think Edelman was right to think it has a role to play in the brain: but working out quite how is another matter. Edelman himself built a novel idea of recategorisation based on the action of re-entrant loops; this part of the theory has not fared very well over the years. The NYT obituary quotes Gunther Stent who once said that as professor of molecular biology and chairman of the neurobiology section of the National Academy of Sciences, he should have understood Edelman's theory - but didn't.

At any rate, we can see that Edelman believed that when a conscious machine was built in the distant future it would be running a selective system of some kind; one that we could well call a machine in everyday terms, though not in the Turing sense. He just might be vindicated one day.

The Turing Test – Is It All Over?

15 June 2014

So, was the brouhaha over the Turing Test justified? It was widely reported last week that the test had been passed for the first time by a chatbot named 'Eugene Goostman'. I think the name itself is a little joke: it sort of means 'well-made ghost man'.

This particular version of the test was not the regular Loebner which we have discussed before (Hugh Loebner must be grinding his teeth in frustration at the apprent ease with which Warwick garnered international media attention), but a session at the Royal Society apparently organised by Kevin Warwick. The bar was set unusually low in this case: to succeed the chatbot only had to convince 30% of the judges that it was human. This was based on the key sentence in the paper by Turing which started the whole thing:

> *I believe that in about fifty years' time it will be possible, to programme computers, with a storage capacity of about 109, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning.*

Fair enough, perhaps, but less impressive than the 50% demanded by other versions; and if 30% is the benchmark, this wasn't actually the first ever pass, because other chatbots like Cleverbot have scored higher in the past. Goostman doesn't seem to be all that special: in an "interview" on BBC radio, with only the softest of questioning, it used one response twice over, word for word.

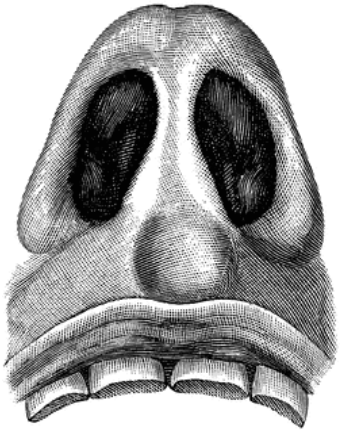
The softness of the questioning does seem to be a crucial variable in Turing tests. If the judges stick to standard small talk and allow the chatbot to take the initiative, quite a reasonable dialogue may result: if the judges are tough it is easy to force the success rate down to zero by various tricks and traps. Iph u spel orl ur werds rong, fr egsampl, uh bott kanot kope, but a human generally manages fine.

Perhaps that wouldn't have worked for Goostman, though because he was presented as relatively ignorant young boy whose first language was not English, giving him some excuse for not understanding things. This stratagem attracted some criticism, but really it is of a piece with chatbot strategy in general; faking and gaming is what it's all about. No-one remotely supposes that Goostman, or Cleverbot, or any of the others, has actually attained consciousness, or is doing anything that could properly be called thinking. Many years ago I believe there were serious efforts to write programs that to some degree imitated the probable mental processes of a human being: they identified a topic, accessed a database of information about it, retained a set of 'attitudes' towards things and tried to construct utterances that made sense in relation to them. It is a weakness of the Turing test that it does not reward that kind of effort; a robot with poor grammar and general knowledge might be readily detectable even though it gave signs of some nascent understanding, while a bot which generates smooth responses without any attempt at understanding has a much better chance of passing.

So perhaps the curtain should be drawn down on the test; not because it has been passed, but because it’s no use.

Smell All About It

28 June 2014



Smell is the most elusive of the senses. Sight is beautifully structured and amenable to analysis in terms of consistent geometry and a coherent domain of colours. Smells… how does one smell relate to another? There just seems to be an infinite number of smells, all one of a kind. We can be completely surprised by an unprecedented smell which is like nothing we ever experienced before, in a way we can't possibly be surprised by a new colour (with some minor possible exceptions). Our olfactory system effortlessly assigns new unique smell experiences to substances that never existed until human beings synthesised them.

There don't even seem to be any words for smells: or at least, the only way we can talk about them is by referring to "the smell of X", as in a "smoky smell" or "the smell of lemons". We don't have to do that to describe shapes or colours: they can be described as "blue", or "square" without our having to say they are "sky-coloured" or "the shape of a box". (Except perhaps in the case of orange? Is "orange" short for 'the colour of oranges'?) Even for taste we have words like "bitter" and "sweet". The only one I can think of for smells is "nidorous', which is pretty obscure - and in order to explain it I have to fall back on saying it describes the "smell of" burning/cooking meat. All we have to describe smells is "strong" and "faint" (my daughter, reading over my shoulder, says what about "pungent"? She does not consider "pungent" to be merely a synonym of "strong" - you may be indifferent to a strong smell, but not to a pungent one, she claims).

With that by way of preamble, let me introduce the interesting question considered here by William Lycan: does smell represent? When we smell, do we smell something? There is a range of possible answers. We might say that when I smell, I smell sausages (for example). Or that I smell a smell (which happens to be the smell of sausages). Or I might say I just have a smell experience: I may know that it's associated with sausage smells and hence with sausages, but in itself it's just an experience.

Lycan (who believes that we smell a gaseous miasma) notes two arguments for something like the last position – that smell doesn't represent anything. First, introspection tells us nothing about what a smell represents. If I were a member of a culture that did not make sausages or eat meat, and had never experienced them, my first nose-full of sausage odour would convey nothing to me beyond itself. It's different for sight: we inherently see things, and when we see our first sausage there can be no doubt we are seeing a thing, even if we do not yet know much about its nature: it would be absurd to maintain we were merely having a visual experience.

The second argument is that smells can't really be wrong: there are no smell illusions. If a car is sprayed with "new car" perfume to make us think that it is fresh off the production line, we may make a mistake about that inference, but our nose was not wrong about the smell, which was real. But representations can always be wrong, so if we can't be wrong, there is no representation.

Lycan is unimpressed by introspective evidence: the mere fact that philosophers disagree about what it tells us is enough, he feels, to discredit it. The second argument fails because it assumes that if smells represent, they must represent their causes: but they might just

represent something in the air. On getting a whiff of my first sausage I would not know what it was, but I might well be moved to say "What's that appetising (or disgusting) smell?" I wouldn't simply say "Golly, I am undergoing a novel olfactory experience for some opaque reason." I think in fact we could go further there and argue that I might well say "What's that I can smell?" – but that doesn't suit Lycan's preferred position. (My daughter intervenes to say "What about 'acrid'?")

Lycan summarises a range of arguments (One is an argument by Richardson that smell is phenomenologically "exteroceptive", inherently about things out there: Lycan endorses this view, but surely relying on phenomenology is smuggling back in the introspection he was so scathing about when the other side invoked it?). His own main argument rests on the view that how something smells is something over and above all the other facts about it. The premise here is very like that in the famous thought experiment of Mary the colour scientist, though Lycan is not drawing the same conclusions at all. He claims instead that:¶

> *I can know the complex of osphresiological fact without knowing how the rose smells because knowing is knowing-under-a-representation… that solution entails that olfactory experience involves representation*.

That does make some sense, I feel (What about "osphresiological"! we're really working on the vocabulary today, aren't we?). You may be asking yourself, however, whether this is a question that needs a single answer. Couldn't we say, yes sometimes smells represent miasmas, but they can also represent sausages; or indeed they can represent nothing.

Lycan, in what I take to be a development of his view, is receptive to the idea of layering: that in fact smells can represent not just a cloud of stuff in the air, but also the thing from which they emanated. That being so I am not completely clear why we should give primacy to the miasma. Two contrary cases suggest themselves. First, suppose there is a odour so faint I don't even perceive it as such consciously, but have a misty sense of salsiccian (alright, I made it up) presence which makes me begin to think about how agreeable a nice Cumberland sausage for lunch might be. Wouldn't we say that in some sense the smell represented sausages to me: but we can't say it represented a miasma because no such thing ever entered my mind?

Second, if we accept layering we might say that the key point is about the essential or the minimal case: we can smell without that smell representing a sausage, but what's the least it can represent and still be a smell? Can it represent nothing? Suppose I dream and have an odd, unrecognisable experience. Later on, when awake, I encounter a Thai curd sausage for the first time and find that the experience I had was in fact an olfactory one, the smell of this particular kind of comestible. My dream experience cannot possibly have represented a sausage, a miasma, a smell, or anything but itself because I didn't know what it was: but, it turns out, it was the smell of curd sausage.

I think your reaction to that is likely to depend on whether you think an experience could be a smell experience without being recognisable as such; if not, you may be inclined to agree with Lycan, who would probably reiterate his view that smells are sensing-under-a-representation. That view entails that there is an ineffability about smell, and Lycan suggests this might help account for the poverty of smell vocabulary that I noted above. Interestingly it turns out that this very point has been attacked by Majid and Burenhult, albeit not in a way that Lycan considers fatal to his case. Majid and Burenhult studied the Jahai, a nomadic hunter-gatherer tribe on the Malaysian peninsula, and found that they have a very rich lexicon of odour terms, such as a

word for “the smell of petrol, smoke and bat droppings” (what, all of them?). It’s just us English speakers, it seems, who are stuck with acrid nidors.

The New Unconsciousness

5 July 2014

Doctors at George Washington found by chance recently that stimulating a patient's claustrum served to disrupt consciousness temporarily (abstract). The patient was being treated for epilepsy, and during this kind of surgery it is normal to use an electrode to stimulate areas of the brain in the target area before surgery to determine their role and help ensure the least possible damage is done to important functions. The claustrum is a sheet-like structure which seems to be well connected to many parts of the brain; Crick and Koch suggested it might be 'the conductor of the orchestra' of consciousness.

New Scientist reported this as the discovery of the 'on/off' switch for consciousness; but that really doesn't seem to be the claustrum's function: there's no reason at the moment to suppose it is involved in falling asleep, or anaesthesia, or other kinds of unconsciousness, The on/off idea seems more like a relatively desperate attempt to explain the discovery in layman's terms, reminiscent of the all-purpose generic tabloid newspaper technology report in Michael Frayn's The Tin Men:

> *British scientists have developed a "magic box", it was learned last night. The new wonder device was tested behind locked doors after years of research. Results were said to have exceeded expectations... ...The device is switched on and off with a switch which works on the same principle as an ordinary domestic light switch...*

Actually, one of the most interesting things about the finding is that the state the patient entered did not resemble sleep or any of those other states; she did not collapse or close her eyes, but instantly stopped reading and became unresponsive - although if she had been asked to perform a repetitive task before stimulation started, she would continue for a few seconds before tailing off. On some occasions she uttered a few incoherent syllables unprompted. This does sound more novel and potentially more interesting than a mere on/off switch. She was unable to report what the experience was like as she had no memory of it afterwards - that squares with the idea that consciousness was entirely absent during stimulation, though it's fair to note that part of her hippocampus, which has an important role in memory formation, had already been removed.

Could Crick and Koch now be vindicated? It seems likely in part: the claustrum seems at least to have some important role - but it's not absolutely clear that it is a co-ordinating one. One of the long-running problems for consciousness has been the binding problem: how the different sensory inputs, processed and delivered at different speeds, somehow come together into a smoothly co-ordinated experience. It could be that the claustrum helps with this, though some further explanation would be needed. As a long shot, it might even be that the claustrum is part of the 'Global Workspace' of the mind hypothesised by Bernard Baars, an idea that is still regularly invoked and quoted.

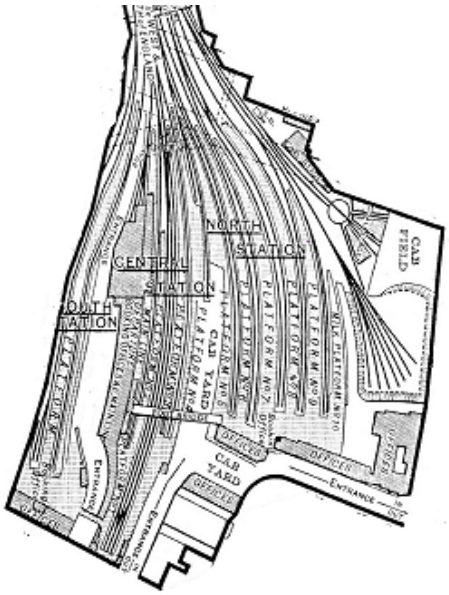
But we must be cautious. All we really know is that stimulating the claustrum disrupted consciousness. That does not mean consciousness happens in the claustrum. If you blow up a major road junction near a car factory, production may cease, but it doesn't mean that the

junction was where the cars were manufactured. Looking at it sceptically we might note that since the claustrum is well connected it might provide an effective way of zapping several important areas at once, and it might be the function of one or more of these other areas that is essential to sustaining consciousness.

However, it is surely noteworthy that a new way of being unconscious should have been discovered. It seems an unprecedentedly pure way, with a very narrow focus on high level activity, and that does suggest that we're close to key functions. It is ethically impossible to put electrodes in anyone's claustrum for mere research reasons, so the study cannot be directly replicated or followed up; but perhaps the advance of technology will provide another way.

Platforms for the Gravy Train

12 July 2014



The European Human Brain Project seems to be running into problems. This Guardian report notes that an open letter of protest has been published by 170 unhappy neuroscientists. They are seeking to influence and extend a review that is due, hoping they can get a change of direction. I don't know a great deal about the relevant EU bureaucracy, but I should think the letter-writers' chances of success are small, not least because in Henry Markram they're up against a project leader who is determined, resourceful, and not lacking support of his own. There's a response to the letter here.

It is a little hard to work out exactly what the disagreement is about; the Guardian seems to smoosh together the current objections of former insiders with the criticisms of those who thought the project was radically premature in the first place. I find myself trying to work out what the protestors want, from Markram's disparaging remarks about them, rather the way we have to reconstruct some ancient heresies from the rebuttals of the authorities, the only place where details survive.

We're told the disagreement is between those who study behaviour at a high level and the project leaders who want to build simulations from the bottom up. In particular some cognitive neuroscience projects have been 'demoted' to partner status. People say the project has been turned into a technology one: Markram says it always was: he suggests that piling up more data is useless and that instead he's doing an ICT project which will provide a platform for integrating the data, and that it's all coming out of an ICT budget anyway.

Us naive outsiders had picked up the impression that the project had a single clear goal; a working simulation of a whole human brain. That is sort of still there but reading the response it seems to be a pretty distant aspiration. Apparently, a mouse brain is going to be done first, but even that is a way off; it's all about the platforms. Earlier documents suggest there will actually be six platforms, only one of which is about brain simulation; the others are neuroinformatics, high performance computing, medical informatics, neuromorphic computing, and neurorobotics - fascinating subjects. The implicit suggestion is that this kind of science can't be done properly just by working in labs and publishing papers, it requires advanced platforms in which research can be integrated. Really? Speaking as a professional bureaucrat myself, I have to say frankly that that sounds uncommonly like the high-grade bollocks emitted by a project leader who has more money than he knows what to do with. The EU in particular is all about establishing unwanted frameworks and common platforms which lie dead in drawers forever after. If people want to share findings, publishing papers is fine (alright, not flawless). If it's about doing actual research, having all the projects captured by a common platform which might embody common errors and common weaknesses doesn't sound like a good idea at all. My brain doesn't know, but my gut says the platforms won't be much use.

Let's be honest, I don't really know what's going on, but if one were cynical one might suppose that the success of the Human Genome Project made the authorities open to other grand

projects, and one on the brain hit the spot. The problem is that we knew what a map of the genome would be like, and we pretty much knew it could be done and how. We don't have a similarly clear idea relating to the brain. However, the concept was appealing enough to attract a big pot of money, both in the EU and then in the US (an even bigger pot). The people who got control of these pots cannot deliver anything like the map of the human genome, but they can buy in the support of fund-hungry researchers by disbursing some of the gold while keeping the politicians and bureaucrats happy by wrapping everything in the afore-mentioned bollocks. The authors of the protest letter perhaps ought to be criticising the whole idea, but really they're just upset about being left out. The deeper sceptics who always said the project was premature - though they may have thought they were talking about brain simulation, not a set of integrative platforms - were probably right; but there's no money in that.

Grand projects like this are probably rarely the best way to control research funding, but they do get funding. Maybe something good somewhere will accidentally get the help it needs; meanwhile we'll be getting some really great European platforms.

Neurons and Free Will

20 July 2014

A few years ago we noted the remarkable research by Fried, Mukamel, and Kreiman which reproduced and confirmed Libet's famous research. Libet, in brief, had found good evidence using EEG that a decision to move was formed about half a second before the subject in question became consciously aware of it; Fried et al produced comparable results by direct measurement of neuron firing.

In the intervening years, electrode technology has improved and should now make it possible to measure multiple sites. The scanty details here indicate that Kreiman, with support from MIT, plans to repeat the research in an enhanced form; in particular he proposes to see whether, having identified the formed intention to move, it is then possible to stop it before the action takes place. This resembles the faculty of 'free won't' by which Libet himself hoped to preserve some trace of free will.

From the MIT article it is evident that Kreiman is a determinist and believes that his research confirms that position. It is generally believed that Libet's findings are incompatible with free will in the sense that they seem to show that consciousness has no effect on our actual behaviour.

That actually sheds an interesting sidelight on our view of what free will is. A decision to move still gets made, after all; why shouldn't it be freely made even though it is unconscious? There's something unsatisfactory about unconscious free will, it seems. Our desire for free will is a desire to be in control, and by that we mean a desire for the entity that does the talking to be in control. We don't really think of the unconscious parts of our mind as being us; or at least not in the same way as that gabby part that claims responsibility for everything (the part of me that is writing this now, for example).

This is a bit odd, because the verbal part of our brain obviously does the verbals; it's strange and unrealistic to think it should also make the decisions, isn't it? Actually if we are careful to distinguish between the making of the decision and being aware of the decision - which we should certainly do, given that one is clearly a first order mental event and the other equally clearly second order - then it ceases to be surprising that the latter should lag behind the former a bit. Something has to have happened before we can be aware of it, after all.

Our unease about this perhaps relates to the intuitive conviction of our own unity. We want the decision and the awareness to be a single event, we want conscious acts to be, as it were, self-illuminating, and it seems to be that that the research ultimately denies us.

It is the case, of course, that the decisions made in the research are rather weird ones. We're not often faced with the task of deciding to move our hands at an arbitrary time for no reason. Perhaps the process is different if we are deciding which stocks and shares to buy? We may think about the pros and cons explicitly, and we can see the process by which the conclusion is

reached; it's not plausible that those decisions are made unconsciously and then simply notified to consciousness, is it?

On the other hand, we don't think, do we, that the process of share-picking is purely verbal? The words flowing through our consciousness are signals of a deeper imaginative modelling, aren't they? If that is the case, then the words might still be lagging. Perhaps the distinction to be drawn is not really between conscious and unconscious, but between simply conscious and explicitly conscious. Perhaps we just shouldn't let the talky bit pretend to be the whole of consciousness just because the rest is silent.

Quanta and Qualia

27 July 2014

Quentin Ruyant has written a thoughtful piece about quantum mechanics and philosophy of mind: in a nutshell he argues both that quantum theory may be relevant to the explanation of consciousness and that consciousness may be relevant to the interpretation of quantum theory.

Is quantum theory relevant to consciousness? Well. of course, some people have said so, notably Sir Roger Penrose and Stuart Hameroff. I think Ruyant is right, though, that the majority of philosophers and probably the majority of physicists dismiss the idea that quantum theory might be needed to explain consciousness. People often suggest that the combination of the two only appeals because both are hard to explain: 'here's one mystery and here's another: maybe one explains the other'. Besides, the brain is far too big and hot and messy for anything other than classical physics to be required

In making the case for the relevance of quantum theory, Ruyant relies on the Hard Problem. His position is that the Hard Problem is not biological but a matter of physics, whereas the Easy Problem, to do with all the scientifically tractable aspects of consciousness, can be dealt with by biology or psychology.

Actually, turning aside from Ruyant's main point, there are some reasons to suggest that quantum physics is relevant to the Easy Problem. Penrose's case, in fact, seems to suggest just that: in his view consciousness is demonstrably non-computable and some kind of novel quantum mechanics is his favoured candidate to fill the gap. Penrose's examples, things like solving mathematical problems, look like 'Easy' Problem matters to me.

Although I don't think anyone advocates the idea, it also seems possible that the 'spooky action at a distance' associated with quantum entanglement might conceivably have something to tell us about intentionality and its remarkable power to address things that are remote and not directly connected with us (I don't think for a moment that that is in fact the case: in fact I don't think quantum physics has anything to do with the Easy Problem: but if we're reviewing the possibilities, I think that would be one.

Anyway, Ruyant is mainly concerned with the Hard Problem, and his argument is that metaphysics and physics are closely related. Topics like the essential nature of physical things straddle the borderline between the two subjects, and it is not at all implausible therefore that the deep physics of quantum mechanics might shed light on the deep metaphysics of phenomenal experience. It seems to me a weakish line of argument, possibly tinged with a bit of prejudice: some physicists are inclined to feel that while their subject deals with the great fundamentals, biology deals only with the chance details of life; sort of a more intellectual kind of butterfly collecting. That kind of thinking is not really well grounded, and its particularly odd to think that biology is irrelevant when considering a phenomenon that, so far as we know, appears only in animals and is definitely linked very strongly with the operation of the brain. John Searle for one argues that 'hard Problem' consciousness arises from natural biological properties of brain tissue. We don't yet know what those properties are, but in his view it's absurd to think that the job of nerves could equally well be performed by beer cans and string. Ruth Millikan,

somewhat differently, has argued that consciousness is purely biological in nature, arising from and defined by evolutionary needs.

I think the truth is that it's difficult to get anywhere at this meta-theoretical level: we don't really decide what kind of theory is most likely to be right, we decide what the true theory most likely is and then root for the kind of theory it happens to be. That, to a great extent, is why quantum theories are not very popular: no-one has come up with one that is cogent and appealling. It seems to me that Ruyant likes the idea of physics-based theories because he favours panpsychism, or panphenomenalism, and so is inclined to think that the essential nature of matter is likely to be the right place to look for a theory.

To be honest, I doubt whether any kind of science can touch the Hard Problem. It's about entities that have no causal properties and are ineffable: how could empirical science ever deal with that? It might well be that a scientist will eventually give us the answer, but it won't be by doing science, because neither calssical nor quantum physics can touch the inexpressible.

Actually, there is a long shot. If Colin McGinn is partly on the right track, it may be that consciousness seems mysterious to us simply because we're not looking at it the right way: our minds won't conceptualise it correctly. Now the same could be true of quantum theory. We struggle with the interpretation of quantum mechanics, but what if we could reorient our brains so that it simply seemed natural, and we groped instead for an acceptable 'interpretation' of spooky classical physics? If we could make such a transformation in our mental orientation, then perhaps consciousness would make sense too? It's possible, but we're back to banging two mysteries together in the hope that some spark will be generated.'

Ruyant's general case, that metaphysicians should be informed by our best physics is hard to argue with. At the moment few philosophers really engage with the physics and few physicists really grasp the philosophy. Why do philosophers avoid quantum physics? Partly, no doubt, just because it's difficult, and relies on mathematics which few philosophers can handle. Partly also, I think there's an unspoken fear that in learning about quantum physics your intuitions will be trained into accepting a particular weltanschauung that might not be helpful. Connected with that is the fear that quantum physics isn't really finished or definitive. Where would I be if I came up with a metaphysical system that perfectly supported quantum theory and then a few years later it turns out that I should have been thinking in terms of string theory? Metaphysicians cross their fingers and hope they can deal with the key issues at a level of generality that means they won't be rudely contradicted by an unexpected advance in physics a few years later.

I suppose what we really need is someone who can come up with a really good specific theory that shows the value of metaphysics informed by physics, but few people are qualified to produce one. I must say that Ruyant seems to be an exception, with an excellent grasp of the theories on both sides of the divide. Perhaps he has a theory of consciousness in his back pocket…?

Strange Smells of the Mind

10 August 2014

An intriguing paper from Benjamin D. Young claims that we can have phenomenal experiences of which we are unaware - although experiences of which we are aware always have phenomenal content. The paper is about smell, though I don't really see why similar considerations shouldn't apply to other senses.

At first sight the idea of phenomenal experience of which we are unaware seems like a contradiction in terms. Phenomenal experience is the subjective aspect of consciousness, isn't it? How could an aspect of consciousness exist without consciousness itself? Young rightly says that it is well established that things we only register subconsciously can affect our behaviour - but that can't include the sort of experience which for some people is the real essence of consciousness, can it?

The only way I can imagine subjectivity going on in my head without me experiencing it is if someone else were experiencing it - not a matter of me experiencing things subconsciously, but of my subconscious being a real separate entity, or perhaps of it all going on in the mind of alternate personality of the kind that seems to occur is Dissociative Identity Disorder (Multiple Personality, as it used to be called).

On further reflection, I don't think that's the kind of thing Young meant at all: I think instead he is drawing a distinction between explicit and inexplicit awareness. So his point is that I can experience qualia without having any accompanying conscious thought about those qualia or the experience.

That's true and an important point. One reason qualia seem so slippery, I think, is that discussion is always in second order terms: we exchange reports of qualia. But because the things themselves are irredeemably first order they have a way of disappearing from the discussion, leaving us talking about their effable accompaniments.

Ironically, something like that may have happened in Young's paper, as he goes on to discuss experiments which allegedly shed light on subjective experience. Smell is a complex phenomenon of course; compared with the neat structure of colours the rambling and apparently inexhaustible structure of smell space is dauntingly hard to grasp. However, smell conveniently has valence in a way that colours don't: some smells are nice and some are nasty. Humans apparently vary their sniff rate partly in response to a smell's valence and Young thinks that this provides an objective, measurable way into the subjectivity of the experience.

Beyond that he goes on to consider mating choice: it seems human beings, like other mammals, choose their mates partly on the basis of smell. I imagine this might be controversial to some, and some of the research Young quotes sounds amusingly naive. In answer to a questionnaire, female subjects rated body odour as an important factor in selecting a sexual partner; well yes, if a guy smells you're maybe not going to date him, huh?

I haven't read the study which was doubtless on a much more sophisticated level, and Young cites a whole wealth of other interesting papers. The problem is that while this is all fascinating

psychologically, none of it can properly bear on the philosophical issue because qualia, the ultimate bearers of subjectivity, are acausal and cannot affect our behaviour. This is shown clearly by the zombie twin argument: my zombie twin has no qualia but his behaviour is ex hypothesi the same as mine.

Still, the use of valence as a way in is interesting. The normal philosophical argument is that we have no way of telling whether my subjective red is your subjective green: but it's hard to argue that m subjective nasty is your subjective nice (unless we also hypothesise that you seek out nasty experiences and avoid nice ones?)

More Qualia Than We Thought?

17 August 2014

What follows are some draft thoughts of my own – Peter

Let's ask a stupid question that may not even be answerable. How many qualia are there? It is generally assumed, I think, that this is like asking how long is a piece of string: that there is an indefinite multiplicity of qualia, that in fact, for every distinguishable sensation there is a matching distinct quale.

As we know, colour is always to the fore in these discussions, and the most common basic example of a quale is probably the colour quale we experience when we see a red rose. I think it is uncontroversial that all sensory experiences come with qualia (uncontroversial among those who believe in qualia at all, that is), although the basis for that appears to be purely empirical; I'm not aware of any arguments to show that all categories of sensory experience must necessarily come with qualia. It would be interesting and perhaps enlightening if some explorers of the phenomenal world reported that, say, the taste of pure water had no accompanying qualia - or that for some, slightly zombish people it had none, while for others it had the full complement of definite phenomenal qualities. To date that has not happened (and perhaps it can't happen?); it seems to be universally agreed that if qualia exist at all, they accompany every sensory experience.

I think it is generally believed that feelings, phenomenal states with no direct relation to details of the external world, have qualia too. Pain qualia are often discussed, with feelings of hunger and pleasure getting occasional mentions; qualia of emotions are also mentioned without provoking controversy. It seems in fact that all experience is generally taken to have accompanying qualia, including dream or hallucinatory experience, and perhaps even certain memories.

In fact there seems to be an interesting, debatable borderline in memory. Vividly recalling a piece of music in real time seems, I would say, to have the same qualia as hearing it live through the ears (Or are the qualia of memories fainter? Do qualia, as a matter of fact, vary in intensity? Or is that idea a kind of contamination from the effable experiences that pair with each quale? It could be so, but then if there is no variation in intensity qualia must be sort of binary, fully on at all times - or fully off - and that doesn't feel quite right either.) In general the same might be claimed for all those memories that involve some 'replay' of experience or feelings; the replay has qualia. Where nothing is held before our attention, on the other hand, there's nothing. The act of merely summoning up a PIN number as we use it does not have its own qualia; there's nothing it is like to recall a password, though there might be something it is like to search the memory for one, and something unpleasant it is like to panic when we fail.

There is certainly room for some phenomenological exploration around these areas, but that more or less exhausts the domain of qualia as I understand it to be generally recognised. I think, however, that it actually stretches a little further than that. There is, in my view, something it is like to be me, something properly ineffable and separable from all the particular sensations and feelings that being me entails. If this is indeed a quale (and of course since this is an ineffable

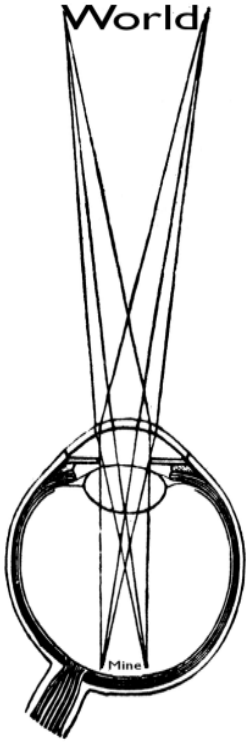
matter I can only appeal to the reader's own introspective research) then I think it's in a category of its own. We might be tempted to assimilate it to the feelings, and say it's the feeling of existing. Or perhaps we might think it's simply the quale that goes with proprioception, the complex but essential sense that tells us where our body is at any moment. Those are respectable qualia no doubt, but I believe there's a quale of being me that goes beyond them.

To that we can add a related and problematic entity which uniquely links the Hard and Easy problems, a phenomenal state we could call the executive quale, that of being in charge. We feel that consciousness is effective, that our conscious decisions have real heft in respect of our behaviour.

This, I think, is the very thing that many people are concerned to deny: the feeling of being causally effective; but to date I don't think it has been regarded as a quale. For some people, who wish to deny both real agency and real subjectivity, the conjunction will seem logical and appealing - to others perhaps less so.

Mine, All Mine

24 August 2014



Following on somewhat from the idea of there being a quale of being me, the latest JCS includes a paper by Marc Slors and Fleur Jongepier about Mineness without Minimal Selves.

'Mineness' here is the quality of our experiences that makes them feel like ours, their first-person givenness. Slors and Jongepier say that the majority of theories explain this in terms of how the experience relates to a minimal self; although different terminology is used all these theories have in common that they rely on 'internal' structure, whereas Slors and Jongepier want instead to advocate a view based on external structure.

What does that all mean? The typical theory - they use Dan Zahavi as a representative case - says that there are three elements; the object experienced, the experiencing, and the subject who experiences. Some have argued that there can't be experience without an experiencer, but we have to remember that a figure as august as Hume held that there was no subject apart from the stream of experience, no core 'me', or at least not one that he could perceive in himself. Now although the subject is indeed not part of the experience per se, it is experientially linked with it in this structure, and that's why it has the feel of belonging to me. In a way this comes down to the commonsensical claim that experiences feel like mine because they relate to me; not surprising that that should be a popular point of view.

That structure, however, takes no account of time: it is, as it were, an instant view: Slors and Jongepier don't think this will do. They quote Metzinger saying that he experiences his leg as having always been part of him, and his experiences as part of a stream of consciousness. They hold that this diachronic aspect of experience cannot be left out. Moreover, while they grant that some version of the internal structure described above could be bodged up to allow for continuous experience, it could not easily take account of more distant memories, which they hold to be equally important.

I'm not sure I see this. We've talked about unfortunate patients who have no ability to form new long or medium term memories: they exist in a kind of small temporal island, never able to remember how they got where they are and hypothesising that they regained consciousness only a few minutes ago. These people are nevertheless perfectly lucid and articulate and apart form the absence of memory seem to be having unimparied experiences which seem to be thier own just as much as anyone else's do. Slors and Jongepier would probably point out that they retain memories from their earlier lives, before their brains were damaged: but if we hypothesise a person with no memories would we also deny them any sense of owning their experiences? I don't really see why.

Anyway, Slors and Jongepier propose a coherentist theory which does not merely say that experience has to fit into a larger 'psychobiography' and dispense with the minimal self. The final, curious element in the theory is the claim that this essential coherence of experience with a background biography is not itself an object of experience. Indeed, it's the fact that the coherence is not experienced that makes the experience feel like mine.

This seems odd at first sight: how can the absence of an experience of coherence make an experience feel like my own? Putting it informally I think the gist is that it is, as it were, the absence of surprise that lets us know things are familiar. Experiences seem like mine because they slide into the stream of consciousness without a splash.

It is an ingenious theory which seems to capture some aspects of phenomenology rather well; but in the end I don't feel motivated to adopt it: it isn't really solving any problems for me. I'm inclined to think that all direct experience seems like mine just because it is direct; my experiences are, as it were, right there, while the external world (and even more so someone else's experiences) are matters of conjecture and inference. I suppose that means I'm hanging on to my minimal self for the moment.

If Atoms Are Real, We're Free

31 August 2014

The Platopus makes a good point about compatibilism (the view that some worthwhile kind of free will is compatible with the standard deterministic account of the world given by physics).

One argument holds that there isn't effectively any difference between compatibilists and those who deny the reality of free will. Both deny that radical (or 'libertarian') free will exists. They agree that there's no magic faculty which interrupts the normal causal process with volitions. Given that level of agreement, isn't it just a matter of what labelling strategy we prefer? Because it's that radical kind of free will that is really at issue: that's what people want, not some watered-down legalistic thing.

That's the argument the Platopus wishes to reject. He accepts that compatibilism involves some redefinition but draws a distinction between illegitimate and legitimate redefinition. As an example of the latter, he proposes the example of atoms. In Greek philosophy, and at first in the modern science which borrowed the word, 'atom' meant something indivisible. There was a period when the atoms of modern physics seemed to be just that, but in due course it emerged that they could in fact be 'split'. One strategy at that point would have been to say, well, it turns out those things were never atoms after all: we must give them a new name and look elsewhere for our indivisible atoms - or perhaps atoms don't actually exist after all. What happened in reality was that we went on calling those particles atoms and gave up our belief that they were indivisible.

In a somewhat similar way, the Platopus argues that it makes sense for us to redefine freedom of the will even though we now know it is not libertarian freedom. The analogy is not perfect, and in some ways the case is actually stronger for free will. Atoms, after all, were originally a hypothesis derived from the purest metaphysics. On one interpretation (just mine, really), the early atomists embraced the idea because they feared that unless the process of division stopped somewhere, the universe would suffer from a radical indeterminism. Division could not stop until the particles were of zero magnitude - non-existent, and how could we make real things out of items which did not themselves exist? They could not have imagined the modern position in which, on one interpretation (yes) as we go more micro the nature of the reality involved changes until the physics has boiled away leaving only maths.

Be that as it was or may be, I think the Platopus is quite right and that the redefinition required by compatibilism is not just respectable but natural and desirable. I think in fact we could go a little further and say that it's not so much a redefinition as a correction of inherent flaws in the pre-theoretical idea of free will.

What do I mean? Well, the original problem here is that the deterministic physical account seems to leave no room for the will. People try to get round that by suggesting different kinds of indeterminism: perhaps we can get something out of chaos theory, or out of quantum mechanics. The problem with those views is that they go too far and typically end up giving us random action: which is no more what we wanted than determined action. Alternatively, old-fashioned libertarians rely on the intervention of the spirit, typically with no satisfactory account

of how the spirit makes decisions or how it manages to intervene. That, I submit, was never really what people meant either: in their Sunday best they might appeal to the action of their soul, but in everyday life having a free choice was something altogether more practical; a matter of not having a knife at your throat.

In short, I'd claim that the pre-theoretical understanding of free will always implicitly took it to be something that went on in a normal physical world, and that's what compatibilism restores, saving the idea from the mad excrescences added by theologians and philosophers.

Myself I think that the kind of indeterminism we can have, and the one we really need, is the one that comes from our power to think about anything. Most processes in the world can be predicted because the range of factors involved can be known and listed to begin with: our mental processes are not like that. Our neurons may work deterministically according to physics, but they allow us to think about anything at any time: about abstractions, remote entities, and even imaginary things. Above all, they allow us somehow to think about the future and enable future contingencies (in some acceptable sense) to influence our present decisions. When our actions are determined by our own thoughts about the future, they can properly be called free.

That is not a complete answer: it defers the mystery of freedom to the mystery of intentionality; but I'll leave that one for now.

Metaconsciousness

6 September 2014

Sci provided some interesting links in his comment on the previous post, one a lecture by Raymond Tallis. Tallis offers some comfort to theists who have difficulty explaining how or why an eternal creator God should be making one-off interventions in the time-bound secular world he had created. Tallis grants that's a bad problem, but suggests atheists face an analogous one in working out how the eternal laws of physics relate to the local and particular world we actually live in.

These are interesting issues which bear on consciousness in at least two important ways, through human agency and the particularity of our experience; but today I want to leave the main road and run off down a dimly-lit alley that looks as if it contains some intriguing premises.

For the theists the problem is partly that God is omniscient and the creator of everything, so whatever happens, he should have foreseen it and arranged matters so that he does not need to intervene. An easy answer is that in fact his supposed interventions are actually part of how he set it up; they look like angry punishments or miraculous salvations to us, but if we could take a step back and see things from his point of view, we'd see it's all part of the eternal plan, set up from the very beginning, and makes perfect sense. More worrying is the point that God is eternal and unchanging; if he doesn't change he can't be conscious. I've mentioned before that our growing understanding of the brain, imperfect as it is, is making it harder to see how God could exist, and so making agnosticism a less comfortable position. We sort of know that human cognition depends on a physical process; how could an immaterial entity even get started? Instead of asking whether God exists, we're getting to a place where we have to ask first how we can give any coherent account of what he could be - and it doesn't look good, unless you're content with a non-conscious God (not necessarily absurd) or a physical old man sitting on a cloud (which to be fair is probably how most Christians saw it until fairly recent times).

So God doesn't change, and our developing understanding is redefining consciousness in ways that make an unchanging consciousness seem to involve a direct contradiction in terms. A changeless process? At this point I imagine an old gentleman dressed in black who has been sitting patiently in the corner, leaning forward with a kindly smile and pointing out that what we're trying to do is understand the mind of God. No mere human being can do that, he says, so no wonder you're getting into a muddle! This is the point where faith must take over.

Well, we don't give up so easily; but perhaps he has a point; perhaps God has another and higher form of consciousness - metaconsciousness, let's say - which resolves all these problems, but in ways we can never really understand. Perhaps when the Singularity comes there will be robots who attain metaconsciousness, too: they may kindly try to explain it to us, but we'll never really be able to get our heads round it.

Now of course, computers can already sail past us in terms of certain kinds of simple capacity: they can remember far more data much more precisely than we can, and they can work quickly through a very large number of alternatives. Even this makes a difference, I've mentioned before that exhaustive analysis by computers has shown that certain chess positions long considered

draws are actually wins for one side: the winning tactics are just so long and complicated that human beings couldn't see them, and can't understand them intuitivel even when they see them played out. But that's not really any help; here we're looking for something much more impressive. What we want to do is take the line which connects an early mammal's level of cognition to ours and extend it until we've gone at least as far beyond the merely human. In facing up to this task, we're rather like Flatlanders trying to understand the third dimension, or ordinary people trying to grasp the fourth: it isn't really possible to get it intuitively, but we ought to be able to say some things about it by extrapolation.

So, early mammal - let's call the beast Em (I don't want to pick a real animal because that will derail us into consideration of how intelligent it really is) - works very largely on an instinctive or stimulus/resp0onse basis. It sees food, it pursues it and attempts to eat it. It lives in a world of concrete and immediate entities and has responses ready for some of them - fairly complex and somewhat variable responses, but fixed in the main. If we could somehow get Em to play chess with us, he would treat his men like a barbarian army, launching them towards us haphazardly en masse or one at a time, and we should have no trouble picking them off.

Human consciousness, by contrast, allows us to consider abstract entities (though we do not well understand their nature), to develop abstract general goals and to make plans and intentions which deal with future and possible events. These plans can also be of great complexity. We can even play out complicated long-range chess strategies if they're not too complex. This kind of thing allows us to do a better job than Em of getting food, though to Em a great deal of our food-related activity is completely opaque and apparently unmotivated. A lot of the time when we're working on activities that will bring us food it will seem to Em as if we're doing nothing, or at any rate, nothing at all related to food.

We can take it, then, that God or a future robot which is metaconscious will have moved on from mere goals to something more sophisticated - metagoals, whatever they are. He, or it, will understand abstractions as well as we understand concrete objects, and will perhaps employ meta-abstractions which they might be a little shakier about. God and the robot will at time have goals, just as we eat food, but their activities in respect of them will be both far more powerful and productive than the simple direct stuff we do and in our eyes utterly unrelated to the simpler goals we can guess at. A lot of the time they may appear to be doing nothing when they are actually pressing forward with an important metaproject.

But look, you may say, we have no reason to suppose this meta stuff exists at all. Em was not capable of abstract thought; we are. That's the end of the sequence; you either got it or you don't got it. We got it: our memory capacity and so on may be improvable, but there isn't any higher realm. Perhaps God's objectives would be longer term and more complex than ours, but that's just a difference of degree.

It could be so, but that's how things would seem to Em, mutatis mutandis. Rocks don't get food, he points out; but we early mammals get it. See food, take food, eat food: we get it. Now humans may see further (nice trick, that hind legs thing) they may get bigger food. But this talk of plans means nothing; there's nothing to your plans and your abstraction thing except getting food. You do get food on a big scale, I notice, but I guess that's really just luck or some kind of magic. Metaconsciousness would seem similarly unimaginable to us, and its results would equally look like magic, or like miracles.

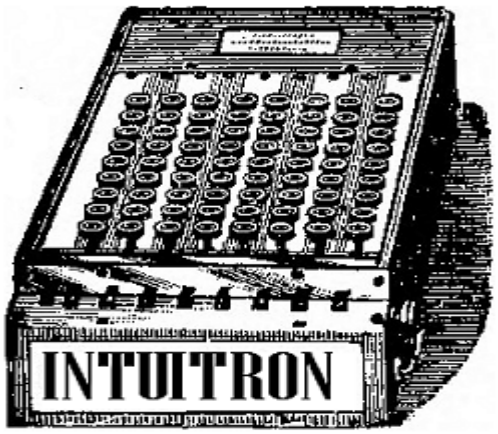
This all fits very well, of course, with Colin McGinn's diagnosis. According to him, there's nothing odd about consciousness in itself, we just lack the mental capacity to deal with it. The mental operations available to us confine us within a certain mental sphere: we are restricted by cognitive closure. It could be that we need metaconsciousness to understand consciousness (and then, unimaginably, metametaconsciousness in order to understand metaconsciousness).

This is an odd place to have ended up, though: we started out with the problem that God is eternal and therefore can't be conscious: if He can't be conscious then He certainly can't ascend to even higher cognitive states, can He? Remember we thought metaconsciousness would probably enable him to understand Platonic abstractions in a way we can't, and even deal with meta-Platonic entities. Perhaps at that level the apparent contradiction between being unchanging and being aware is removed or bypassed, rather the way that putting five squares together in two dimensions is absurd but a breeze in three: hell, put six together and make a cube of it!

Do I really believe in metaconsciousness? No, but excuse me; I have to go and get food.

The Intuitional Problem

14 September 2014



Mark O'Brien gives a good statement of the computationalist case here; clear, brief, and commendably mild and dispassionate. My impression is that enthusiasm for computationalism - approximately, the belief that human thought is essentially computational in nature - is not what it was. It's not that computationalists lost the argument, it's more that the robots failed to come through. What AI research delivered has so far been, in this respect, much less than the optimists had hoped.

Anyway O'Brien's case rests on two assumptions:¶

- Naturalism is true.
- The laws of physics are computable, that is, any physical process can be simulated to any desired degree of precision by a computer.

It's immediately clear where he's going. T0 represent it crudely, the intuition here is that naturalism means the world ultimately consists of physical processes, any physical process can run on a computer, ergo anything in the world can run on a computer, ergo it must be possible to run consciousness on a computer.

There's an awful lot packed into those two assumptions. O'Brien tackles one issue with the idea of simulation: namely that simulating something isn't doing it for real. A simulated rainstorm doesn't make us wet. His answer is that simulation doesn't produce physical realities, but it does seem to work for abstract things. I think this is basically right. If we simulate a flight to Paris, we don't end up there; but the route calculated by the program is the actual route; it makes no sense to say it's only a simulated route, because it's actually identical with the one we should use if we really went to Paris. So the power of simulation is greater for informational entities than for physical ones, and it's not unreasonable to suggest that consciousness seems more like a matter of information than of material stuff.

There's a deeper point, though. To simulate is not to reproduce: a simulation is the reproduction of the relevant aspects of the thing simulated. It's implied that some features of the thing simulated are left out, ones that don't matter. That's why we get the different results for our Parisian exercise: the simulation necessarily leaves our actual physical locations untouched; those are irrelevant when it comes to describing the route, but essential when it comes to actually visiting Paris.

The problem is we don't know which properties are relevant to consciousness, and to assume they are the kind of thing handled by computation simply begs the question. It can't be assumed without an argument that physical properties are irrelevant here: John Searle and Roger Penrose in different ways both assert that they are of the essence. Even if consciousness doesn't rely quite so brutally as that on the physical nature of the brain, we need to start with a knowledge of how consciousness works. Otherwise, we can't tell whether we've got the right properties in our simulation - *even if they are in principle computational*.

I don't myself think Searle or Penrose are right: but I think it's quite likely that the causal relationships in cognitive processes are the kind of essential thing a simulation would have to incorporate. This is a serious problem because there are reasons to think computer simulations never reproduce the causal relationships intact. In my brain event A causes event B and that's all there is to it: in a computer, there's always a script involved. At its worst what we get is a program that holds up flag A to represent event A and then flag B to represent event B: but the causality is mediated through the program. It seems to me this might well be a real issue.

O'Brien tackles another of Searle's arguments: that you can't get semantics from syntax: ie, you can't deal with meanings just by manipulating digits. O'Brien's strategy here is to assume a robot that behaves pretty much the way I do: does it have beliefs? It says it does, and it behaves as if it did. Perhaps we're not willing to concede that those are real beliefs: OK, let's call them beliefs*. On examination it turns out that the differences between beliefs and beliefs* are nugatory: so on gorunds of parsimony if nothing else we should assume they are the same.

The snag here is that there are no robots that behave the way I do. We've had sixty years of failure since Turing: you can't just have it as an assumption that our robot pals are self-evidently achievable (alas). We know that human beings, when they do translation for example, extract meanings and then put the meanings into other words, whereas the most successful translation programs avoid meanings altogether and simply swap text strings for text strings according to a kind of mighty look-up table.

That kind of strategy won't work when dealing with the formless complexity of the real world: you run into the analogues of the Frame Problem or you just don't get really started. It doesn't even work that well for language: we know now that human understanding of language relies on pragmatic Gricean implicatures, and no-one can formalise those.

Finally O'Brien turns to qualia, and here I agree with him on the broad picture. He describes some of the severe difficulties around qualia and says, rightly I think, that in the end it comes down to competing intuitions. All the arguments for qualia are essentially thought experiments: if we want, we can just say 'no' to all of them (as Dennett and the Churchlands, for example, do). O'Brien makes a kind of zombie argument: my zombie twin, who lacks qualia but resembles me in all other respects, would claim to have qualia and would talk about them just the way we do. So the explanation for talk about qualia is not qualia themselves: given that, there's no reason to think we ourselves have them.

Up to a point: but we get the conclusion that my zombie twin talks about qualia purely ex hypothesi: it's just specified. It's not an explanation, and I think that's what we really need to be in a position to dismiss the strong introspective sense most people have that qualia exist. If we could actually explain what makes the Twin talk about qualia, we'd be in a much better position.

So I mostly disagree, but I salute O'Brien's exposition, which is really helpful.

Theorobotics

22 September 2014

A recent piece on Salon notes how a robot is giving lectures in theology - or perhaps it would be more accurate to say that it's being used as a prop for some theology lectures. It helps dramatise certain human issues, either the 'strong' ones about it lacking the immortal soul human beings are taken to have in Christian thought, or some 'weak' ones about more general ethical issues.

Nothing wrong with that; in fact I've heard it argued that all thinking robots would be theists, because to them it would seem obvious, almost self-evident, that conscious entities need a creator. No doubt D.A.V.I.D helps to raise interest, but he doesn't seem half as provocative as the Jesus automaton described here; not a modern robot but a feature of the medieval church robot scene, apparently a far livelier business than we could ever have guessed.

It's certainly true that those old automata had a deep impact on Western thought about the mind. Descartes describes hydraulic ones, and it's clear that they helped form his idea of the human body as a mere machine. The study of anatomy was backing this up - Leonardo da Vinci, for example, had already concluded on the basis of anatomy alone that the brain was the centre from which the body was controlled. Together these two influences banished older ideas of volition acting throughout the body, with your arm moving because you just wanted it to, impelled by your unintermediated volition. These days, of course, some actually think we have gone too far with our brain-centrism, and need to bring in ideas of embodiment and mind extension; but rightly or wrongly the automata undoubtedly changed our minds dramatically.

The same kind of thing happened when effective computers came on the scene. Before then it had seemed obvious that though the body might be a machine, the mind categorically was not; now there was a persuasive case for thinking our minds as well as our bodies might be machines, and I think our idea of consciousness has been reshaped gradually since so that it can fill the role of 'the thing machines can't do' for those who think there is such a thing.

It might be that this has distorted our way of looking at consciousness, which never occupied an important place in ancient thought, and does not really feature in the same way in non-western traditions (at least so far as I can tell). So perhaps robots shouldn't be teaching us about the mind. On the other hand, they sometimes come up with interesting stuff. Dennett's discussion of the frame problem is a nice example. Most people take the frame problem - in essence, dealing with all the small background details of real-world situations which multiply indefinitely, are probably irrelevant, but might just come back to bite you - as a problem for AI: but Dennett thoughtfully suggested that it was in fact a problem for all forms of intelligence. It was just that the human brain dealt with it so smoothly we'd never noticed it before: but to explain how the brain dealt with it was at least as problematic as building a robot that could handle it. In this way the robots had given us a new insight into human cognition. So perhaps we should listen to them?'

Sam Harris the Mystic

29 September 2014

I must admit to not being very familiar with Sam Harris' work: to me he's been primarily a member of the Four Horsemen of New Atheism: Dawkins, Dennett, Hitchens and that other one... However in the video here he expresses a couple of interesting views, one about the irreducible subjectivity of consciousness, the other about the illusory nature of the self. His most recent book - first chapter here - apparently seeks to reinterpret spirituality for atheists; he seems basically to be a rationalist Buddhist (there is of course no contradiction involved in becoming a Buddhist while remaining an atheist).

It's a slight surprise to find an atheist champion who does not want to do away with subjectivity. Harris accepts that there is an interior subjective experience which cannot simply be reduced to its objective, material correlates: he likens the two aspects to two faces of a coin. If you like, you can restrict yourself to talking about one face of the coin, but you can't go on to say that the other doesn't really exist, or that features of the heads side are really just features of the tails side seen from another angle. So far as it goes, this is all very sensible, and I think the majority of people would go along with it quite a way. What's a little odd is that Harris seems content to rest there: it's just a feature of the world that it has these two aspects, end of story. Most of us still want some explanation: if not a reduction then at least some metaphysics which allows us to go on calling ourselves monists in a respectable manner.

Harris' second point is also one that many others would agree with, but not me. The self, he says, is an illusion: there is no consistent core which amounts to a self. In these arguments I feel the sceptics are often guilty of knocking down straw men: they set up a ridiculous version of the self and demolish that without considering more reasonable ideas. So, they deny that that there is any specific part of the brain that can be identified with the self, they deny the existence of a Cartesian Theatre, or they deny any unchanging core. But whoever said the self had to be unchanging or simple?

Actually, we can give a pretty good account of the self without ever going beyond common sense. Human beings are animals, which means I'm a big live lump of meat which has a recognisable identity at the simple physical and biological level: to deny that takes a more radical kind of metaphysical scepticism than most would be willing to go in for. The behaviour of that ape-like lump of meat is also governed by a reasonably consistent set of memories and beliefs. That's all we need for a self, no mystery here, folks, move along please.

Now of course my memories and my beliefs change, as does the size and shape of the beast they inhabit. At 56 I'm not the same as I was at 6. But so what? Rivers, as Heraclitus reminds us, never contain exactly the same water at two different moments: they rise and fall, they meander and change course. We don't have big difficulties over believing in rivers, though, or even in the Nile or the Amazon in particular. There may be doubts about what to treat as the true origin of the Nile, but people don't go round saying it's an illusion (unless they've gone in for some of that more radical metaphysics). On this, I think Dennett's conception of the self as a centre of

narrative gravity is closer to the truth than most, though he has, unfairly, I think, been criticised for equivocating over its reality.

Frequently what people really want to deny is not the self so much as the soul. Often they also want to deny that there is a special inward dimension: but as we've seen, Harris affirms that. He seems instead almost to be denying the qualic sense of self I suggested a while back. Curiously, he thinks that we can in fact, overcome the illusion of selfhood: in certain exalted states he thinks we can transcend ourselves and see the world as it really is.

This is strange, because you would expect the illusion of self to stem from something fundamental about conscious experience (some terrible bottleneck, some inherent limitation), not from small, adjustable details of chemistry. Can selfhood really be a mental disease caused by an ayahuasca deficiency? Harris asserts that in these exalted states we're seeing the world as it truly is, but wouldn't that be the best state for us to stay in? You'd think we would have evolved that way if seeing reality just needed some small tweaks to the brain.

It does look to me as if Harris' thinking has been conditioned a little too much by Buddhism. He speaks with great respect of the rational basis of Buddhism, pointing out that it requires you to believe its tenets merely because they can be shown to be true, whereas Christianity seems to require as an arbitrarily specified condition of salvation your belief in things that are acknowledged to be incredible. I have a lot of sympathy for that point of view; but the snag is that if you rely on reasoning your reasoning has to be watertight: and, at the risk of giving offence, Buddhism's isn't.

Buddhism tells us that the world is in constant change; that change inevitably involves pain, and that to avoid the pain we should avoid the world. As it happens, it adds, the mutable world and the selves we think inhabit it are mere illusions, so if we can dispel those illusions we're good.

But that's a bit of a one-sided outlook, surely? Change can also involve pleasure, and in most of our lives there's probably a neutral to positive balance; so surely it makes more sense to engage and try to improve that balance than opt out? Moreover, should we seek to avoid pain? Perhaps we ought to endure it, or even seek it out? Of course, people do avoid pain, but why should we give priority to the human aversion to pain and ignore the equally strong human aversion to non-existence? And what does it mean to call the whole world an illusion: isn't an illusion precisely something that isn't really part of the world? Everything we see is smoke and mirrors, very well, but aren't smoke and mirrors (and indeed, tears, as Virgil reminds us) things?

A sceptical friend once suggested to me that all the major religions were made up by frustrated old men, often monks: the one thing they all agree on is that being a contemplative like them is just so much more worthwhile than you'd think at first sight, and that the cheerful ordinary life they missed out on is really completely worthless if not sinful. That's not altogether accurate - Taoism, for example, praises ordinary life (albeit with a bit of a smirk on its lips); but it does seem to me that Buddhism is keener to be done with the world than reason alone can justify.

It must be said that Harris' point of view is refreshingly sophisticated and nuanced in comparison to the Dawkinsian weltanschauung; he seems to have the rare but valuable ability to apply his critical faculties to his own beliefs as well as other people's. I really should read some of his books.

Quantized Qualia

5 October 2014

You've heard of splitting the atom: W. Alex Escobar wants to split the quale. His recent paper proposes that in order to understand subjective experience we may need to break it down into millions of tiny units of experience. He proposes a neurological model which to my naive eyes seems reasonable: the extraordinary part is really the phenomenology.

Like a lot of qualia theorists Escobar seems to have based his view squarely on visual experience, and the idea of micro-qualia is perhaps inspired by the idea of pixels in digitised images, or other analytical image-handling techniques. Why would the idea help explain qualia?

I don't think Escobar explains this very directly, at least from a philosophical point of view, but you can see why the idea might appeal to some people. Panexperientialists, for example, take the view that there are tiny bits of experience everywhere, so the idea that our minds assemble complex experiences out of micro-qualia might be quite congenial to them. As we know, Christof Koch says that consciousness arises from the integration of information, so perhaps he would see Escobar's theory as offering a potentially reasonable phenomenal view of the same process.

Unfortunately Escobar has taken a wrong turning, as others have done before, and isn't really talking about ineffable qualia at all: instead, we might say he is merely effing the effable.

Ineffability, the quality of being inexpressible, is a defining characteristic of qualia as canonically understood in the philosophical literature. I cannot express to you what redness is like to me; if I could, you would be able to tell whether it was the same as your experience. If qualia could be expressed, my zombie twin (who has none) would presumably become aware of their absence; when asked what it was like to see red, he would look puzzled and admit he didn't really know, whereas ex hypothesi he gives the same fluent and lucidly illuminating answers that I do - in spite of not having the things we're both talking about.

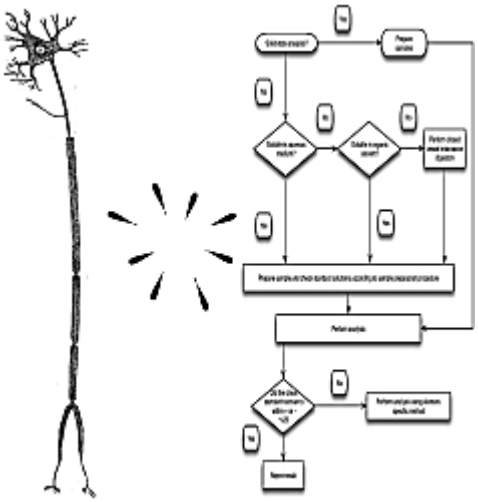
Qualia, in fact, have no causal effects and cannot be part of the scientific story. That doesn't mean Escobar's science is wrong or uninteresting, just that what he's calling qualia aren't really the philosophically slippery items of experience we keep chasing in vain in our quest for consciousness.

Alright, but setting that aside, is it possible that real qualia could be made up of many micro-qualia? No, it absolutely isn't! In physics, a table can seem to be a single thing but actually be millions of molecules. Similarly, what looks like a flat expanse of uniform colour on a screen may actually be thousands of pixels. But qualia are units of experience; what they seem like is what they are. They don't seem like a cloud of micro-qualia, and so they aren't. Now there could be some neuronal or psychological story going on at a lower level which did involve micro units; but that wouldn't make qualia themselves splittable. What they are like is all there is to them; they can't have a hidden nature.

Alas, Escobar could not have noticed that, because he was too busy effing the Effable.

A Different Gap

12 October 2014



We're often told that when facing philosophical problems, we should try to 'carve them at the joints'. The biggest joint on offer in the case of consciousness has seemed to be the 'explanatory gap' between the physical activity of neurons and the subjective experience of consciousness. Now, in the latest JCS, Reggia, Monner, and Sylvester suggest that there is another gap, and one where our attention should rightly be focussed.

They suggest that while the simulation of certain cognitive processes has proceeded quite well, the project of actually instantiating consciousness computationally has essentially got nowhere. That project, as they say, is affected by a variety of problems about defining and recognising success. But the real problem lies in an unrecognised issue: the computational explanatory gap. Whereas the explanatory gap we're used to is between mind and brain, the computational gap is between high-level computational algorithms and low-level neuronal activity. At the high level, working top-down, we've done relatively well in elucidating how certain kinds of problem-solving, goal-directed kinds of computation work, and been able to simulate them relatively effectively. At the neuronal, bottom-up level we've been able to explain certain kinds of pattern-recognition and routine learning systems. The two different kinds of processing have complementary strengths and weaknesses, but at the moment we have no clear idea of how one is built out of the other. This is the computational explanatory gap.

In philosophy, the authors plausibly claim, this important gap has been overlooked because in philosophical terms these are all 'easy problem' matters, and so tend to be dismissed as essentially similar matters of detail. In psychology, by contrast, the gap is salient but not clearly recognised as such: the lower-level processes correspond well to those identified as sub-conscious, while the higher-level ones match up with the reportable processes generally identified as conscious.

If Reggia, Monner and Sylvester are right, the well-established quest for the neural correlates of consciousness has been all very well, but what we really need is to bridge the gap by finding the computational correlates of consciousness. As a project, bridging this gap looks relatively promising, because it is all computational. We do not need to address any spooky phenomenology, we do not need to wonder how to express 'what it is like', or deal with anything ineffable; we just need to find the read-across between neural networking and the high-level algorithms which we can sort of see in operation. That may not be easy, but compared to the Hard Problem it sounds quite tractable. If solved, it will deliver a coherent account right from neural activity through to high-level decision making.

Of course, that leaves us with the Hard Problem unsolved, but the authors are optimistic that success might still banish the problem. They draw an analogy with artificial life: once it seemed obvious that there was a mysterious quality of being alive, and it was unclear how simple chemistry and physics could ever account for it. That problem of life has never been solved in

terms, but as our understanding of the immensely subtle chemistry of living things has improved, it has gradually come to seem les and less obvious that it is really a problem. In a similar way the authors hope that if the computational explanatory gap can be bridged, so that we gradually develop a full account of cognitive processes from the ground-level firing of neurons up to high-level conscious problem-solving, the Hard Problem will gradually cease to seem like something we need to worry about.

That is optimistic, but not unreasonably so, and I think the new perspective offered is a very interesting and plausible one. I'm convinced that the gap exists and that it needs to be bridged: but I'm less sure that it can easily be done. It might be that Reggia, Monner, and Sylvester are affected in a different way by the same kind of outlook they criticise in philosophers: these are all computational problems, so they're all tractable. I'm not sure how we can best address the gap, and I suspect it's there not just because people have failed to recognise it, but because it is also genuinely difficult to deal with.

For one thing, the authors tend to assume the problem is computational. It's not clear that computation is of the essence here. The low-level processes at a neuronal level don't look to be based on running any algorithm - that's part of the nature of the gap. High-level processes may be capable of simulation algorithmically, but that doesn't mean that's the way the brain actually does it. Take the example of catching a ball - how do we get to the right place to intercept a ball flying through the air? One way to do this would be some complex calculations about perspective and vectors: the brain could abstract the data, do the sums, and send back the instructions that result. We could simulate that process in a computer quite well. But we know - I think - that that isn't the way it's actually done: the brain uses a simpler and quicker process which never involves abstract calculation, but is based on straight matching of two inputs; a process which incidentally corresponds to a sub-optimal algorithm, but one that is good enough in practice. We just run forward if the elevation of the ball is reducing and back if it's increasing. Fielders are incapable of predicting where a ball is going, but they can run towards the spot in such a way as to be there when the ball arrives. It might be that all the 'higher-level' processes are like this, and that an attempt to match up with ideally-modelled algorithms is therefore categorically off-track.

Even if those doubts are right, however, it doesn't mean that the proposed re-framing of the investigation is wrong or unhelpful, and in fact I'm inclined to think it's a very useful new perspective.

Are We Really Conscious?

19 October 2014

Yes: I feel pretty sure that anyone reading this is indeed conscious. However, the NYT recently ran a short piece from Michael S. A. Graziano which apparently questioned it. A fuller account of his thinking is in this paper from 2011; the same ideas were developed at greater length in his book Consciousness and the Social Brain

I think the startling headline on the NYT piece misrepresents Graziano somewhat. The core of his theory is that awareness is in some sense a delusion, the reality of which is simple attention. We have ways of recognising the attention of other organisms, and what it is fixed on (the practical value of that skill in environments where human beings may be either hunters or hunted is obvious): awareness is just our garbled version of attention. he offers the analogy of colour. The reality out there is different wavelengths of light: colour, our version of that, is a slightly messed-up, neatened, artificial version which is nevertheless very vivid to us in spite of being artificial to a remarkably large extent.

I don't think Graziano is even denying that awareness exists, in some sense: as a phenomenon of some kind it surely does: what he means is more that it isn't veridical: what it tells us about itself, and what it tells us about attention, isn't really true. As he acknowledges in the paper, there are labelling issues here, and I believe it would be possible to agree with the substance of what he says while recasting it in terms that look superficially much more conventional.

Another labelling issue may lurk around the concept of attention. On some accounts, it actually presupposes consciousness: to direct one's attention towards something is precisely to bring it to the centre of your consciousness. That clearly isn't what Graziano means: he has in mind a much more basic meaning. Attention for him is something simple like having your sensory organs locked on to a particular target. This needs to be clear and unambiguous, because otherwise we can immediately see potential problems over having to concede that cameras or other simple machines are capable of attention; but I'm happy to concede that we could probably put together some kind of criterion, perhaps neurological, that would fit the bill well enough and give Graziano the unproblematic materialist conception of attention that he needs.

All that looks reasonably OK as applied to other people, but Graziano wants the same system to supply our own mistaken impression of awareness. Just as we track the attention of others with the false surrogate of awareness, we pick up our own attentive states and make the same kind of mistake. This seems odd: when I sense my own awareness of something, it doesn't feel like a deduction I've made from objective evidence about my own behaviour: I just sense it. I think Graziano actually wants it to be like that for other people too. He isn't talking about rational, Sherlock Holmes style reasoning about the awareness of other people, he has in mind something like a deep, old, lizard-brain kind of thing; like the sense of somebody there that makes the hairs rise on the back of the neck and your eyes quickly saccade towards the presumed person.

That is quite a useful insight, because what Graziano is concerned to deny is the reality of subjective experience, of qualia, in a word. To do so he needs to be able to explain why awareness seems so special when the reality is nothing more than information processing. I think this remains a weak spot in the theory, but the idea that it comes from a very basic system whose whole function is to generate a feeling of 'something there' helps quite a bit, and is at least moderately compatible with my own intuitions and introspections. What Graziano really relies on is the suggestion that awareness is a second-order matter: it's a cognitive state about other cognitive states, something we attribute to ourselves and not, as it seems to be, directly about the real world. It just happens to be a somewhat mistaken cognitive state.

That still leaves us in some difficulty over the difference between me and other people. If my sense of my own awareness is generated in exactly the same way as my sense of the awareness of others, it ought to seem equally distant - but it doesn't, it seems markedly more present and far less deniable.

More fundamentally, I still don't really see why my attention should be misperceived. In the case of colours, the misrepresentation of reality comes from two sources, I think. One is the inadequacy of our eyes; our brain has to make do with very limited data on colour (and on distance and other factors) and so has to do things like hypothesising yellow light where it should be recognising both red and green, for example. Second, the brain wants to make it simple for us and so tries desperately to ensure that the same objects always look the same colour, although the wavelengths being received actually vary according to conditions. I find it hard to see what comparable difficulties affect our perception of attention. Why doesn't it just seem like attention? Graziano's view of it as a second-order matter explains how it can be wrong about attention, but not really why.

So I think the theory is less radical than it seems, and doesn't quite nail the matter on some important points: but it does make certain kinds of sense and at the very least helps keep us roused from our dogmatic slumbers. Here's a wild thought inspired (but certainly not endorsed) by Graziano. Suppose our sense of qualia really does come from a kind of primitive attention-detecting module. It detects our own attention and supplies that qualic feel, but since it also (in fact primarily) detects other people's attention, should it not also provide a bit of a qualic feel for other people too? Normally when we think of our beliefs about other people, we remain in the explicit, higher realms of cognition: but what if we stay at a sort of visceral level, what if we stick with that hair-on-the back-of the-neck sensation? Could it be that now and then we get a whiff of other people's qualia? Surely too heterodox an idea to contemplate.

Perceptronium

26 October 2014

Earlier this year Tononi's Integrated Information Theory (IIT) gained a prestigious supporter in Max Tegmark, professor of Physics at MIT. The boost for the theory came not just from Tegmark's prestige, however; there was also a suggestion that the IIT dovetailed neatly with some deep problems of physics, providing a neat possible solution and the kind of bridge between neuroscience, physics and consciousness that we could hardly have dared to hope for.

Tegmark's paper presents the idea rather strangely, suggesting that consciousness might be another state of matter like the states of being a gas, a liquid, or solid. That surely can't be true in any simple literal sense because those particular states are normally considered to be mutually exclusive: becoming a gas means ceasing to be a liquid. If consciousness were another member of that exclusive set it would mean that becoming conscious involved ceasing to be solid (or liquid, or gas), which is strange indeed. Moreover Tegmark goes on to name the new state 'perceptronium' as if it were a new element. He clearly means something slightly different, although the misleading claim perhaps garners him sensational headlines which wouldn't be available if he were merely saying that consciousness arose from certain kinds of subtle informational organisation, which is closer to what he really means.

A better analogy might be the many different forms carbon can take according to the arrangement of its atoms: graphite, diamond, charcoal, graphene, and so on; it can have quite different physical properties without ceasing to be carbon. Tegmark is drawing on the idea of computronium proposed by Toffoli and Margolus. Computronium is a hypothetical substance whose atoms are arranged in such a way that it consists of many tiny modules capable of performing computations. There is, I think, a bit of a hierarchy going on here: we start by thinking about the ability of substances to contain information; the ability of a particular atomic matrix to encode binary information is a relatively rigorous and unproblematic idea in information theory. Computronium is a big step up from that: we're no longer thinking about a material's ability to encode binary digits, but the far more complex functional property of adequately instantiating a universal Turing machine. There are an awful lot of problems packed into that 'adequately'.

The leap from information to computation is as nothing, however, compared to the leap apparently required to go from computronium to perceptronium. Perceptronium embodies the property of consciousness, which may not be computational at all and of which there is no agreed definition. To say that raises a few problems is rather an understatement.

Aha! But this is surely where the IIT comes in. If Tononi is right, then there is in fact a hard-edged definition of consciousness available: it's simply integrated information, and we can even say that the quantity required is Phi. We can detect it and measure it and if we do, perceptronium becomes mathematically tractable and clearly defined. I suppose if we were curmudgeons we might say that this is actually a hit against the IIT: if it makes something as absurd as

perceptronium a possibility, there must be something pretty wrong with it. We're surely not that curmudgeonly, but there is something oddly non-dynamic here. We think of consciousness, surely, as a process, a function: but it seems we might integrate quite a lot of information and simply have it sit there as perceptronium in crystalline stillness; the theory says it would be conscious, but it wouldn't do anything. We could get round that by embracing the possibility of static conscious states, like one frame out of the movie film of experience; but Tegmark, if I understand him right, adds another requirement for consciousness: autonomy, which requires both dynamics and independence; so there has to be active information processing, and it has to be isolated from outside influence, much the way we typically think of computation.

The really exciting part, however, is the potential linkage with deep cosmological problems - in particular the quantum factorisation problem. This is way beyond my understanding, and the pages of equations Tegmark offers are no help, but the gist appears to be that quantum mechanics offers us a range of possible universes. If we want to get 'physics from scratch', all we have to work with is, in Tegmark's words,¶

> *two Hermitian matrices, the density matrix p encoding the state of our world and the Hamiltonian H determining its time-evolution...*

Please don't ask me to explain; the point is that the three things don't pin down a single universe; there are an infinite number of acceptable solutions to the equations. If we want to know why we've got the universe we have - and in particular why we've got classical physics, more or less, and a world with an object hierarchy - we need something more. Very briefly, I take Tegmark's suggestion to be that consciousness, with its property of autonomy, tends naturally to pick out versions of the universe in which there are similarly integrated and independent entities - in other words the kind of object-hierarchical world we do in fact see around us. To put it another way and rather baldly, the universe looks like this because it's the only kind of universe which is compatible with the existence of conscious entities capable of perceiving it.

That's some pretty neat footwork, although frankly I have to let Tegmark take the steering wheel through the physics and in at least one place I felt a little nervous about his driving. It's not a key point, but consider this passage:¶

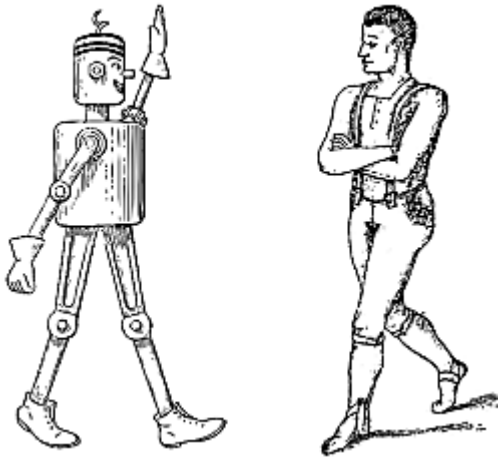
> *Indeed, Penrose and others have speculated that gravity is crucial for a proper understanding of quantum mechanics even on small scales relevant to brains and laboratory experiments, and that it causes non-unitary wavefunction collapse. Yet the Occam's razor approach is clearly the commonly held view that neither relativistic, gravitational nor non-unitary effects are central to understanding consciousness or how conscious observers perceive their immediate surroundings: astronauts appear to still perceive themselves in a semi-classical 3D space even when they are effectively in a zero-gravity environment, seemingly independently of relativistic effects, Planck-scale spacetime fluctuations, black hole evaporation, cosmic expansion of astronomically distant regions, etc*

Yeah... no. It's not really possible that a professor of physics at MIT thinks that astronauts float around their capsules because the force of gravity is literally absent, is it? That kind of 'zero g' is just an effect of being in orbit. Penrose definitely wasn't talking about the gravitational effects of the Earth, by the way; he explicitly suggests an imaginary location at the centre of the Earth so that they can be ruled out. But I must surely be misunderstanding.

So far as consciousness is concerned, the appeal of Tegmark's views will naturally be tied to whether one finds the IIT attractive, though they surely add a bit of weight to that idea. So far as quantum factorisation is concerned, I think he could have his result without the IIT if he wanted: although the IIT makes it particularly neat, it's more the concept of autonomy he relies on, and that would very likely still be available even if our view of consciousness were ultimately somewhat different. The linkage with cosmological metaphysics is certainly appealing, essentially a sensible version of the Anthropic Principle which Stephen Hawking for one has been prepared to invoke in a much less attractive form.

Inscrutable Robots

2 November 2014



Petros Gelepithis has A Novel View of Consciousness in the International Journal of Machine Consciousness (alas, I can't find a freely accessible version). Computers, as such, can't be conscious, he thinks, but robots can; however, proper robot consciousness will necessarily be very unlike human consciousness in a way that implies some barriers to understanding.

Gelepithis draws on the theory of mind he developed in earlier papers, his theory of noèmona species. (I believe he uses the word noèmona mainly to avoid the varied and potentially confusing implications that attach to mind-related vocabulary in English.) It's not really possible to do justice to the theory here, but it is briefly described in the following set of definitions, an edited version of the ones Gelepithis gives in the paper.¶

1. Definition 1. For a human H, a neural formation N is a structure of interacting sub-cellular components (synapses, glial structures, etc) across nerve cells able to influence the survival or reproduction of H.
2. Definition 2. For a human, H, a neural formation is meaningful (symbol Nm), if and only if it is an N that influences the attention of that H.
3. Definition 3. The meaning of a novel stimulus in context (Sc), for the human H at time t, is whatever Nm is created by the interaction of Sc and H.
4. Definition 4. The meaning of a previously encountered Sc, for H is the prevailed Np of Np
5. Definition 5. H is conscious of an external Sc if and only if, there are Nm structures that correspond to Sc and these structures are activated by H's attention at that time.
6. Definition 6. H is conscious of an internal Sc if and only if the Nm structures identified with the internal Sc are activated by H's attention at that time.
7. Definition 7. H is reflectively conscious of an internal Sc if and only if the Nm structures identified with the internal Sc are activated by H's attention and they have already been modified by H's thinking processes activated by primary consciousness at least once.

For Gelepithis consciousness is not an abstraction, of the kind that can be handled satisfactorily by formal and computational systems. Instead it is rooted in biology in a way that very broadly recalls Ruth Millikan's views. It's about attention and how it is directed, but meaning comes out of the experience and recollection of events related to evolutionary survival.

For him this implies a strong distinction between four different kinds of consciousness; animal consciousness, human consciousness, machine consciousness and robot consciousness. For machines, running a formal system, the primitives and the meanings are simply inserted by the human designer; with robots it may be different. Through, as I take it, living a simple robot life they may, if suitably endowed, gradually develop their own primitives and meanings and so attain their own form of consciousness. But there's a snag…¶

> *Robots may be able to develop their own robot primitives and subsequently develop robot understanding. But no robot can ever understand human meanings; they can only interact successfully with humans on the basis of processing whatever human-based primitives and other notions were given…*

Different robot experience gives rise to a different form of consciousness. They may also develop free will. Human beings act freely when their Acquired Belief and Knowledge (ABK) over-rides environmental and inherited influences in determining their behaviour; robots can do the same if they acquire an Own Robot Cognitive Architecture, the relevant counterpart. However, again…¶

> *A future possible conscious robotic species will not be able to communicate, except on exclusively formal bases, with the then Homo species.*

'then Homo' because Gelepithis thinks it's possible that human predecessors to Homo Sapiens would also have had distinct forms of consciousness (and presumably would have suffered similar communication issues).

Now we all have slightly different experiences and heritage, so Gelepithis' views might imply that each of our consciousnesses is different. I suppose he believes that intra-species commonality is sufficient to make those differences relatively unimportant, but there should still be some small variation, which is an intriguing thought.

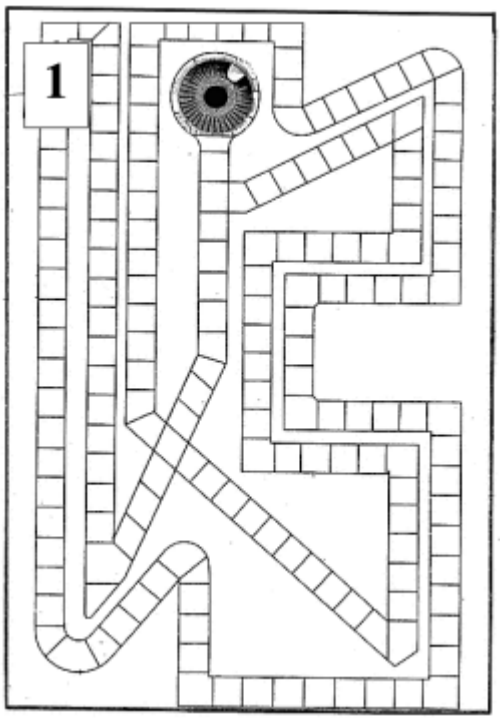
As an empirical matter, we actually manage to communicate rather well with some other species. Dogs don't have our special language abilities and they don't share our lineage or experiences to any great degree; yet very good practical understandings are often in place. Perhaps it would be worse with robots, who would not be products of evolution, would not eat or reproduce, and so on. Yet it seems strange to think that as a result their actual consciousness would be radically different?

Gelepithis' system is based on attention, and robots would surely have a version of that; robot bodies would no doubt be very different from human ones, but surely the basics of proprioception, locomotion, manipulation and motivation would have to have some commonality?

I'm inclined to think we need to draw a further distinction here between the form and content of consciousness. It's likely that robot consciousness would function differently from ours in certain ways: it might run faster, it might have access to superior memory, it might, who knows, be multi-threaded. Those would all be significant differences which might well impede communication. The robot's basic drives might be very different from ours: uninterested in food, sex, and possibly even in survival, it might speak lyrically of the joys of electricity which must remain ever hidden from human beings. However, the basic contents of its mind would surely be of the same kind as the contents of our consciousness (hallo, yes, no, gimme, come here, go away) and expressible in the same languages?

Intellectual Catastrophe

10 November 2014



Scott has a nice discussion of our post-intentional future (or really our non-intentional present, if you like) here on Scientia Salon. He quotes Fodor saying that the loss of 'common-sense intentional psychology' would be the greatest intellectual catastrophe ever: hard to disagree, yet that seems to be just what faces us if we fully embrace materialism about the brain and its consequences. Scott, of course, has been exploring this territory for some time, both with his Blind Brain Theory and his unforgettable novel Neuropath; a tough read, not because the writing is bad but because it's all too vividly good.

Why do we suppose that human beings uniquely stand outside the basic account of physics, with real agency, free will, intentions and all the rest of it? Surely we just know that we do have intentions? We can be wrong about what's in the world; that table may be an illusion; but our intentions are directly present to our minds in a way that means we can't be wrong about them - aren't they?

That kind of privileged access is what Scott questions. Cast your mind back, he says, to the days before philosophy of mind clouded your horizons, when we all lived the unexamined life. Back to Square One, as it were: did your ignorance of your own mental processes trouble you then? No: there was no obvious gaping hole in our mental lives; we're not bothered by things we're not aware of. Alas, we may think we've got a more sophisticated grasp of our cognitive life these days, but in fact the same problem remains. There's still no good reason to think we enjoy an epistemic privilege in respect of our own version of our minds.

Of course, our understanding of intentions works in practice. All that really gets us, though, is that it seems to be a viable heuristic. We don't actually have the underlying causal account we need to justify it; all we do is apply our intentional cognition to intentional cognition…¶

> *it can never tell us what cognition is simply because solving that problem requires the very information intentional cognition has evolved to do without.*

Maybe then, we should turn aside from philosophy and hope that cognitive science will restore to us what physical science seems to take away? Alas, it turns out that according to cognitive science our idea of ourselves is badly out of kilter, the product of a mixed-up bunch of confabulation, misremembering, and chronically limited awareness. We don't make decisions, we observe them, our reasons are not the ones we recognise, and our awareness of our own mental processes is narrow and error-filled.

That last part about the testimony of science is hard to disagree with; my experience has been that the more one reads about recent research the worse our self-knowledge seems to get.

If it's really that bad, what would a post-intentional world look like? Well, probably like nothing really, because without our intentional thought we'd presumably have an outlook like that of dogs, and dogs don't have any view of the mind. Thinking like dogs, of course, has a long and respectable philosophical pedigree going back to the original Cynics, whose name implies a d0g-level outlook. Diogenes himself did his best to lead a doggish, pre-intentional life, living rough, splendidly telling Alexander the Great to fuck off and less splendidly, masturbating in public ('Hey, I wish I could cure hunger too just by rubbing my stomach'). Let's hope that's not where we're heading.

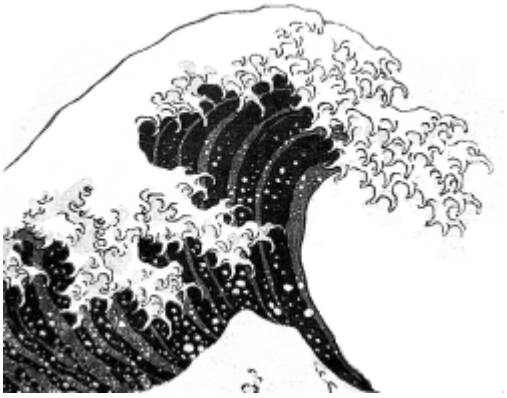
However, that does sort of indicate the first point we might offer. Even Diogenes couldn't really live like a dog: he couldn't resist the chance to make Plato look a fool or hold back when a good zinger came to mind. We don't really cling to our intentional thoughts because we believe ourselves to have privileged access (though we may well believe that); we cling to them because believing we own those thoughts in some sense is just the precondition of addressing the issue at all, or perhaps even of having any articulate thoughts about anything. How could we stop? Some kind of spontaneous self-induced dissociative syndrome? Intensive meditation? There isn't really any option but to go on thinking of our selves and our thoughts in more or less the way we do, even if we conclude that we have no real warrant for doing so.

Secondly, we might suggest that although our thoughts about our own cognition are not veridical, that doesn't mean our thoughts or our cognition don't exist. What they say about the contents of our mind is wrong perhaps, but what they imply about there being contents (inscrutable as they may be) can still be right. We don't have to be able to think correctly about what we're thinking in order to think. False ideas about our thoughts are still embodied in thoughts of some kind.

Is 'Keep Calm and Carry On' the best we can do?

A Third Wave?

16 November 2014



An article in the Chronicle of Higher Education (via the always-excellent Mind Hacks) argues cogently that as a new torrent of data about the brain looms, we need to ensure that it is balanced by a corresponding development in theory. That must surely be right: but I wonder whether the torrent of new information is going to bring about another change in paradigm, as the advent of computers in the twentieth century surely did?

We have mentioned before the two giant projects which aim to map and even simulate the neural structure of the brain, one in America, one in Europe. Other projects elsewhere and steady advances in technology seem to indicate that the progress of empirical neuroscience, already impressive, is likely to accelerate massively in coming years.

The paper points out that at present, in spite of enormous advances, we still know relatively little about the varied types of neurons and what they do; and much of what we think we do know is vague, tentative, and possibly misleading. Soon, however, 'there will be exabytes (billions of gigabytes) of data, detailing what vast numbers of neurons do, in real time'.

The authors rightly suggest that data alone is no good without theoretical insights: they fear that at present there may be structural issues which lead to pure experimental work being funded while theory, in spite of being cheaper, is neglected or has to tag along as best it can. The study of the mind is an exceptionally interdisciplinary business, and they justifiably say research needs to welcome 'mathematicians, engineers, computer scientists, cognitive psychologists, and anthropologists into the fold'. No philosophers in the list, I notice, although the authors quote Ned Block approvingly. (Certainly no novelists, although if we're studying consciousness the greatest corpus of exploratory material is arguably in literature rather than science. Perhaps that's asking a bit too much at this stage: grants are not going to be given to allow neurologists to read Henry as well as William James, amusing though that might be.)

I wonder if we're about to see a big sea change; a Third Wave? There's no doubt in my mind that the arrival of practical computers in the twentieth century had a vast intellectual impact. Until then philosophy of mind had not paid all that much attention to consciousness. Free Will, of course, had been debated for centuries, and personal identity was also a regular topic; but consciousness per se and qualia in particular did not seem to be that important until - I think - the seventies or eighties when a wide range of people began to have actual experience of computers. Locke was perhaps the first person to set out a version of the inverted spectrum argument, in which the blue in your mind is the same as the yellow in mine, and vice versa; but far from its being a key issue he mentions it only to dismiss it: we all call the same real world colours by the same names, so it's a matter of no importance. Qualia? Of no philosophical interest.

I think the thing is that until computers actually appeared it was easy to assume, like Leibniz, that they could only be like mills: turning wheels, moving parts, nothing there that resembles a mind. When people could actually see a computer producing its results, they realised that there

was actually the same kind of incomprehensible spookiness about it as there was in the case of human thought; maybe not exactly the same mystery, but a pseudo-magic quality far above the readily-comprehensible functioning of a mill. As a result, human thought no longer looked so unique and we needed something to stand in as the criterion which separated machines from people. Our concept of consciousness got reshaped and promoted to play that role, and a Second Wave of thought about the mind rolled in, making qualia and anything else that seemed uniquely human of special concern.

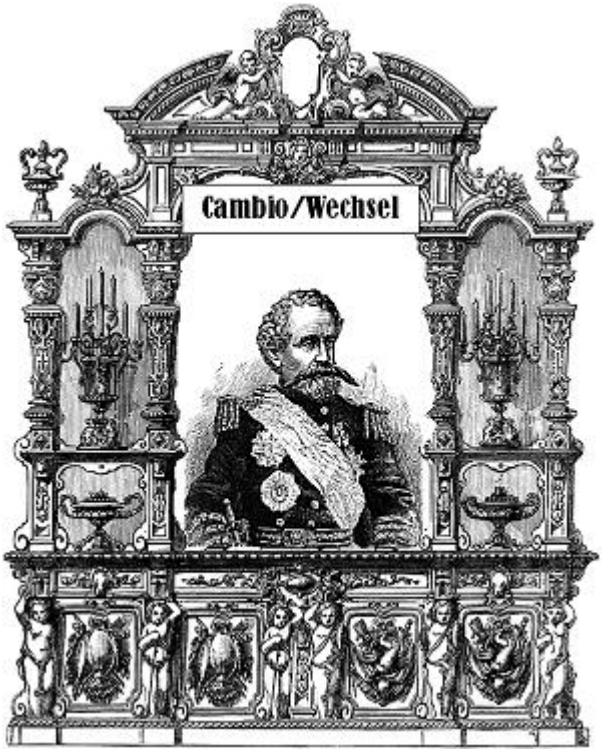
That wave included another change, though, more subtle but very important. In the past, the answer to questions about the mind had clearly been a matter of philosophy, or psychology; at any rate an academic issue. We were looking for a heavy tome containing a theory. Once computers came along, it turned out that we might be looking for a robot instead. The issues became a matter of technology, not pure theory. The unexpected result was that new issues revealed themselves and came to the fore. The descriptive theories of the past were all very well, but now we realised that if we wanted to make a conscious machine, they didn't offer much help. A good example appears in Dan Dennett's paper on cognitive wheels, which sets out a version of the Frame Problem. Dennett describes the problem, and then points out that although it is a problem for robots, it's just as mysterious for human cognition; actually a deep problem about the human mind which had never been discussed; it's just that until we tried to build robots we never noticed it. Most philosophical theories still have this quality, I'm afraid, even Dennett's: OK, so I'm here with my soldering iron or my keyboard: how do I make a machine that adopts the intentional stance? No clue.

For the last sixty years or so I should say that the project of artificial intelligence has set the agenda and provided new illumination in this kind of way. Now it may be that neurology is at last about to inherit the throne. If so, what new transformations can we expect? First I would think that the old-fashioned computational robots are likely to fall back further and that simulations, probably using neural network approaches, are likely to come to the fore. Grand Union theories, which provide coherent accounts from genetics through neurology to behaviour, are going to become more common, and build a bridgehead for evolutionary theories to make more of an impact on ideas about consciousness. However, a lot of things we thought we knew about neurons are going to turn out to be wrong, and there will be new things we never spotted that will change the way we think about the brain. I would place a small bet that the idea of the connectome will look dusty and irrelevant within a few years, and that it will turn out that neurons don't work quite the way we thought.

Above all though, the tide will surely turn for consciousness. Since about 1950 the game has been about showing what, if anything, was different about human beings; why they were not just machines (or why they were), and what was unique about human consciousness. In the coming decade I think it will all be about how consciousness is really the same as many other mental processes. Consciousness may begin to seem less important, or at any rate it may increasingly be seen as on a continuum with the brain activity of other animals; really just a special case of the perfectly normal faculty of… Well, I don't actually know what, but I look forward to finding out.

Personhood Week

23 November 2014



Personhood Week, at National Geographic, is a nice set of short pieces briefly touring the issues around the crucial but controversial issue of what constitutes a person.

You won't be too surprised to hear that in my view personhood is really all about consciousness. The core concept for me is that a person is a source of intentions - intentions in the ordinary everyday sense rather than in the fancy philosophical sense of intentionality (though that too). A person is an actual or potential agent, an entity that seeks to bring about deliberate outcomes. There seems to be a bit of a spectrum here; at the lower level it looks as if some animals have thoughtful and intentional behaviour of the kind that would qualify them for a kind of entry-level personhood. At its most explicit, personhood implies the ability to articulate complicated contracts and undertake sophisticated responsibilities: this is near enough the legal conception. The law, of course, extends the idea of a person beyond mere human beings, allowing a form of personhood to corporate entities, which are able to make binding agreements, own property, and even suffer criminal liability. Legal persons of this kind are obviously not 'real' ones in some sense, and I think the distinction corresponds with the philosophical distinction between original (or intrinsic, if we're bold) and derived intentionality. The latter distinction comes into play mainly when dealing with meaning. Books and pictures are about things, they have meanings and therefore intentionality, but their meaningfulness is derived: it comes only from the intentions of the people who interpret them, whether their creators or their 'audience'. My thoughts, by contrast, really just mean things, all on their own and however anyone interprets them: their intentionality is original or intrinsic.

So, at least, most people would say (though others would energetically contest that description). In a similar way my personhood is real or intrinsic: I just am a person; whereas the First Central Bank of Ruritania has legal personhood only because we have all agreed to treat it that way. Nevertheless, the personhood of the Ruritanian Bank is real (hypothetically, anyway; I know Ruritania does not exist - work with me on this), unlike that of, say, the car Basil Fawlty thrashed with a stick, which is merely imaginary and not legally enforceable.

Some, I said, would contest that picture: they might argue that ;a source of intentions makes no sense because 'people' are not really sources of anything; that we are all part of the universal causal matrix and nothing comes of nothing. Really, they would say, our own intentions are just the same as those of Banca Prima Centrale Ruritaniae; it's just that ours are more complex and reflexive - but the fact that we're deeming ourselves to be people doesn't make it any the less a matter of deeming. I don't think that's quite right - just because intentions don't feature in

physics doesn't mean they aren't rational and definable entities - but in any case it surely isn't a hit against my definition of personhood; it just means there aren't really any people.

Wait a minute, though. Suppose Mr X suffers a terrible brain injury which leaves him incapable of forming any intentions (whether this is actually possible is an interesting question: there are some examples of people with problems that seem like this; but let's just help ourselves to the hypothesis for the time being). He is otherwise fine: he does what he's told and if supervised can lead a relatively normal-seeming life. He retains all his memories, he can feel normal sensations, he can report what he's experienced, he just never plans or wants anything. Would such a man no longer be a person?

I think we are reluctant to say so because we feel that, contrary to what I suggested above, agency isn't really necessary, only conscious experience. We might have to say that Mr X loses his legal personhood in some senses; we might no longer hold him responsible or accept his signature as binding, rather in the way that we would do for a young child: but he would surely retain the right to be treated decently, and to kill or injure him would be the same crime as if committed against anyone else. Are we tempted to say that there are really two grades of personhood that happen to coincide in human beings, a kind of 'Easy Problem' agent personhood on the one hand and a 'Hard Problem' patient personhood? I'm tempted, but the consequences look severely unattractive. Two different criteria for personhood would imply that I'm a person in two different ways simultaneously, but if personhood is anything, it ought to be single, shouldn't it? Intuitively and introspectively it seems that way. I'd feel a lot happier if I could convince myself that the two criteria cannot be separated, that Mr X is not really possible.

What about Robot X? Robot X has no intentions of his own and he also has no feelings. He can take in data, but his sensory system is pretty simple and we can be pretty sure that we haven't accidentally created a qualia-experiencing machine. He has no desires of his own, not even a wish to serve, or avoid harming human beings, or anything like that. Left to himself he remains stationary indefinitely but given instructions he does what he's told: and if spoken to, he passes the Turing Test with flying colours. In fact, if we ask him to sit down and talk to us, he is more than capable of debating his own personhood, showing intelligence, insight, and understanding at approximately human levels. Is he a person? Would we hesitate over switching him off or sending him to the junk yard?

Perhaps I'm cheating. Robot X can talk to us intelligently, which implies that he can deal with meanings. If he can deal with meanings, he must have intentionality, and if he has that perhaps he must, contrary to what I said, be able to form intentions after all - so perhaps the conditions I stipulated aren't possible after all? And then, how does he generate intentions, as a matter of fact? I don't know, but on one theory intentionality is rooted in desires or biological drives. The experience of hunger is just primally about food, and from that kind of primitive aboutness all the fancier kinds are built up. Notice that it's the experience of hunger, so arguably if you had no feelings you couldn't get started on intentionality either! If all that is right, neither Robot X nor Mr X is really as feasible as they might seem: but it still seems a bit worrying to me..

Early Qualia

1 December 2014

The problem of qualia is in itself a very old one, but it is expressed in new terms. My impression is that the actual word 'qualia' only began to be widely used (as a hot new concept) in the 1970s. The question of whether the colours you experience in your mind are the same as the ones I experience in mine, on the other hand, goes back a long way. I'm not aware of any ancient discussions, though I should not be at all surprised to hear that there is one in, say, Sextus Empiricus (if you know one please mention it): I think the first serious philosophical exposition of the issue is Locke's in the Essay Concerning Human Understanding:¶

> *"Neither would it carry any imputation of falsehood to our simple ideas, if by the different structure of our organs, it were so ordered, that the same object should produce in several men's minds different ideas at the same time; e.g. If the idea, that a violet produced in one man's mind by his eyes, were the same that a marigold produces in another man's, and vice versa. For since this could never be known: because one man's mind could not pass into another man's body, to perceive, what appearances were produced by those organs; neither the ideas hereby, nor the names, would be at all confounded, or any falsehood be in either. For all things, that had the texture of a violet, producing constantly the idea, which he called blue, and those that had the texture of a marigold, producing constantly the idea, which he as constantly called yellow, whatever those appearances were in his mind; he would be able as regularly to distinguish things for his use by those appearances, and understand, and signify those distinctions, marked by the names blue and yellow, as if the appearances, or ideas in his mind, received from those two flowers, were exactly the same, with the ideas in other men's minds."*

Interestingly, Locke chose colours which are (near enough) opposites on the spectrum; this inverted spectrum form of the case has been highly popular in recent decades. It's remarkable that Locke put the problem in this sophisticated form; he managed to leap to a twentieth-century outlook from a standing start, in a way. It's also surprising that he got in so early: he was, after all, writing less than twenty years after the idea of the spectrum was first put forward by Isaac Newton. It's not surprising that Locke should know about the spectrum; he was an enthusiastic supporter of Newton's ideas, and somewhat distressed by his own inability to follow them in the original. Newton, no courter of popularity, deliberately expressed his theories in terms that were hard for the layman, and scientifically speaking, that's what Locke was. Alas, it seems the gap between science and philosophy was already apparent even before science had properly achieved a separate existence: Newton would still have called himself a natural philosopher, I think, not a scientist.

It's hard to be completely sure that Locke did deliberately pick colours that were opposite on the spectrum – he doesn't say so, or call attention to their opposition (there might even be some room for debate about whether 'blue' and 'yellow are really opposite) but it does seem at least that he felt that strongly contrasting colours provided a good example, and in that respect at least he anticipated many future discussions. The reason so many modern theorists like the idea is that they believe a switch of non-opposite colour qualia would be detectable, because the spectrum would no longer be coherent, while inverting the whole thing preserves all the relationships intact and so leaves the change undetectable. Myself, I think this argument is a mistake, inadvertently transferring to qualia the spectral structure which actually belongs to the objective counterparts of colour qualia. The qualia themselves have to be completely indistinguishable, so it doesn't matter whether we replace yellow qualia with violet or orange ones, or for that matter, with the quale of the smell of violets.

Strangely enough though Locke was not really interested in the problem; on the contrary, he set it out only because he was seeking to dismiss it as an irrelevance. His aim, in context, was to argue that simple perceptions cannot be wrong, and the possibility of inconsistent colour judgements - one person seeing blue where another saw yellow - seemed to provide a potential counter-argument which he needed to eliminate. If one person sees red where another sees green, surely at least one of them must be wrong? Locke's strategy was to admit that different people might have different ideas for the same percept (nowadays we would probably refer to these subjective ideas of percepts as qualia), but to argue that it doesn't matter because they will always agree about which colour is, in fact yellow, so it can't properly be said that their ideas are wrong. Locke, we can say, was implicitly arguing that qualia are not worth worrying about, even for philosophical purposes.

This 'so what?' line of thought is still perfectly tenable. We could argue that two people looking at the same rose will not only agree that it is red, but also concur that they are both experiencing red qualia; so the fact that inwardly their experiences might differ is literally of no significance - obviously of no practical significance, but arguably also metaphysically nugatory. I don't know of anyone who espouses this disengaged kind of scepticism, though; more normally people who think qualia don't matter go on to argue that they don't exist, either. Perhaps the importance we attach to the issue is a sign of how our attitudes to consciousness have changed: it was itself a matter of no great importance or interest to Locke. I believe consciousness acquired new importance with the advent of serious computers, when it became necessary to find some quality with which we could differentiate ourselves from machines. Subjective experience fit the bill nicely.

Actual Consciousness

7 December 2014

Ted Honderich's latest work *Actual Consciousness* is a massive volume. He has always been partial to advancing his argument through a comprehensive review (and rejection) of every other opinion on the subject in question. Here, that approach produces a hefty book which in practice is about the whole field of philosophy of consciousness. There is a useful video here at IAI of Ted grumpily failing to summarise the whole theory in the allotted time and confessing with the same alarming frankness that characterised his autobiography to wanting to be as famous as Bach or Chomsky, and not thinking he was going to be. If you want to see the whole thing you'll have to sign up (free); but they do have a number of good discussions of consciousness.

The theory Honderich is advancing is a further version of the externalism which we discussed a while ago; that for you to be conscious is in some sense for something to exist (or to be real, hence the 'actual' in Actual Consciousness). At first sight this thesis has always seemed opaque to the point of wilful obscurity, and the simplest readings seem to make it either vacuous (for you to be conscious is for a state of consciousness to exist) or just evidently wrong (for you to be conscious is for the object of your awareness to exist). He means - he must mean - something subtler than that, and a few more clues can only be welcome.

First though, we survey the alternatives. Honderich suggests (and few would disagree) that the study of consciousness has been vastly complicated by differing or inadequate definitions. This has led philosophers to talk past each other or work themselves into confusions. Above all, Honderich thinks virtually everyone has at some point fallen into circularity, smuggling into their definitions terms that already include consciousness in one form or another.

He sets out five leading ideas: these are not actually the five parts into which he would carve consciousness himself (he would analyse it into three: perceptual, cognitive and affective consciousness) but these are the ideas he feels we need to address. They are: qualia, 'something it is like for a thing to be that thing', subjectivity, intentionality, and phenomenality. More normally these days we divide the field in two initially, and at first glance four of Honderich's views look like the same thing from different angles. When there is 'something it is like', that's the phenomenal aspect of experience as had by a subject and characterised by the presence of qualia. But let's not be hasty.

Having reviewed briefly what various people have said about qualia, Honderich notes that one thing seems clear; that it is always conceived of as distinct from, and hence accompanied by, another form of consciousness. Some people certainly assert that qualia are the real essence of consciousness or at any rate of the interesting part of it; but it does seem to be true that no-one proposes conscious states that include qualia and nothing else. That in itself doesn't amount to circularity, though.

The next leading idea is *something it is like* to see red, or whatever. Nagel's phrase is unhelpful but somehow powerfully persuasive. We all sort of know what it is getting at. Honderich notes that Nagel himself offered an improved version that leaves out the suggestion that a

comparison is going on; to be conscious, in this version, is for there to be something that is how it is for you (to see red or whatever). What does this all really mean? Honderich suspects that it comes down to there being something it is like, or something that is how it is, for you to be conscious (of something red, eg), once again a case of circularity. I don't really see it; it seems to me that Nagel offers an equation; consciousness is there being something it is like; Honderich pounces: Aha! But there being something it is like is being conscious! That just seems to be travelling back from the second term of the equation to the first, not showing that the the second term requires or contains the first. I'm simplifying rather a lot, so perhaps I've missed something. But so far as I can see while Honderich justly complains that the formula is uninformative, the only circularity is one he inserted himself.

Subjectivity for Honderich means the existence of a subject. The word, as he acknowledges, can often be used as more or less a synonym for one of the two senses already discussed: in fact I should say that that is the standard meaning. But it's true that consciousness is tied up with the notion of an experiencing self or subject (and those who deny the existence of one are often sceptical about the other). Honderich suggests that it is implicit in the idea of a subject that the subject is conscious, and though we can raise quibbles over sleeping or anaesthetised subjects, he is surely on firmer ground in seeing circularity here. To define consciousness in terms of a subject is circular because to be a subject you have to be conscious. But nobody does that, or at least, no-one I can think of. It's sort of accepted that you need to have your consciousness sorted out before you can have your conscious agent.

With *intentionality* we come on to something distinctly different; this is the quality of aboutness or directedness singled out by Brentano. Honderich bases his comments on Brentano's account, which he quotes at full length. It's only fair to note in passing that Brentano was not talking about consciousness; rather he asserted that intentionality was the mark of the mental; but there is obviously a connection and we might well try to argue that intentionality was constitutive of consciousness.

Honderich notes an intractable problem over the objects of intentionality; they don't have to exist. We can think about imaginary or even absurd things just as easily as about real ones. But if we are not thinking about a real slipper when we think of Cinderella's glass one, then surely we're not really thinking about the actual one the dog is chewing in the corner, either; perhaps the real objects are just our ideas or images of the slipper, or whatever. If we don't take that path, suggests Honderich, then this intentionality business is no great help; if we do suppose that thinking about things is thinking about a mental image, then we're back with circularity because it would have to be a conscious image, wouldn't it?

Would it? I'm not totally sure it would; wouldn't the theory be that it becomes a conscious image when it's an object of conscious thought, but not otherwise or in itself? Otherwise we seem to have a weird doubling up going on. But anyway, it's too clear, in my opinion, that thinking about a thing is thinking about that thing, not thinking about an idea of it; we have to find some other way round the problem of non-existent objects of thought. So we're left with the complaint that intentionality does not explain consciousness - and that's true enough; it's at least as much a part of the problem.

With *phenomenality* we're back with a word that could be taken as meaning much the same as subjectivity, and referring to the same stuff as qualia or something it is like. Honderich draws on Ned Block's much-cited distinction between access or a-consciousness and phenomenal or p-consciousness, and attacks David Chalmers for having said that qualia, subjectivity,

phenomenality and so on, are essentially different ways of talking about the same thing. I'm with Chalmers; yes, the different terms embody different ways of approaching the problem, but there's little doubt that if we could crack one, translating the solution into terms of the other approaches would be relatively trivial. Oddly, instead of recapitulating the claimed important distinctions, which woulod seem the natira;l thing to do at this point, Honderich seems to argue that if Chalmers thinks all these things are the same thing, they must in fact all be examples of something more fundamental, in which case why doesn't Chalmers talk about the fundamental thing?¶

> *If there are the grammatical or subtle differences between the terms for the phenomena, the things do make up 'approximately the same class of phenomena'. What class is that? To speak differently, what are these things examples of ? In fact didn't we have to have some idea of that in order to bring the examples together in the first place? What brings the different things together? What is this general fact of consciousness? It has to exist, doesn't it? Chalmers, I'd say, has credit for bringing the things together, but he might have asked about the general fact, mightn't he?*

This is strange; to assert that a number of terms refer to the same thing is not necessarily to assert that they are all in fact yet another thing. My best guess is that Honderich wants to manoeuvre Chalmers into seeming circularity again, but if so I don't think it comes off.

Honderich goes on to an extended review of the territory and what others have said, but I propose to cut to the chase. Cutting to the chase, by the way, is something Honderich himself is very bad at, or rather pathologically averse to. He has a style all his own, characterised by a constant backing away from saying anything directly; he prefers to mention a few things one might say in relation to this matter, some indeed that others have at times suggested he himself might not always have been committed to denial of, that is to say these considerations might be ones - not by any means to characterise exhaustively, but nevertheless to bring forward as previously hinted - perhaps to be felt to be most significantly indicated or at any rate we might choose, not yet to think, but to entertain the possibility, of considering as such. Sometimes you really want to kick him.

Anyway, to get to the point: Honderich is an externalist; he thinks your perception of x is something that happens out there where x is real and physical, not in your head. There is an extension to this to take care of those cases where we think about things that are imaginary or otherwise non-physical; in such cases the same thing is going on, it's just that the objects perceived are representations in our mind. In a sense this is externalism simply redirected to objects that happen to be internal. Of course, how anything comes to be a mental representation is itself a non-trivial issue.

Honderich says that for you to be conscious of something is for something to be real, to exist. This puzzling or vacuous-seeming formula is underpinned by the eyebrow-raising idea of subjective physicality. This is like a kind of Philosopher's Stone; it means that what we perceive can be both actual and physical in a perfectly normal way, yet subjective in the way required by consciousness. How can we possibly eat our cake this way and yet still have it? It's kind of axiomatic that the actual qualities of physical things don't depend on the observer (yes, I know in modern physics that's a can of worms, but not one we need to open here), while subjective qualities absolutely do; my subjective impressions may be quite different to yours.

How is this trick to be pulled off?¶

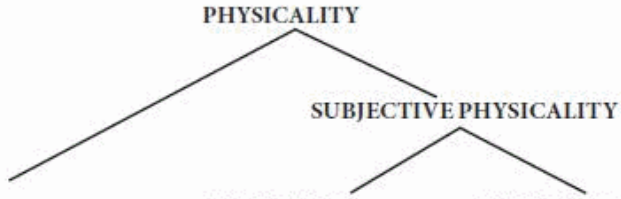
> *The general answer to the question of what is actual with your perceptual consciousness, putting aside that in which it may issue immediately, is a part, piece or stage of a subjective physical world of several dependencies, out there in space, and nothing else whatever. Your being conscious now is exactly and nothing more than this severally-dependent fact external to you of a room's existing…*

It looks at first sight as if this talk of worlds may be the answer. The room exists subjectively for you and also physically, but in a world of its own, or of your own, which would explain how it can be different from the subjective experience of others; they have different worlds. This would be alarming, however, because it suggests a proliferation of worlds whose relationships would be problematic and whose ontology profligate. We're talking some serious and bizarre metaphysics, of which there is no sign elsewhere.

I don't think that is at all what Honderich has in mind. Instead I think we need to remember that he doesn't mean by subjectivity what everyone else means. He just means there is a subject. So his description of consciousness comes down to this; that there is a real object of consciousness, whether in the world or in the brain, which is unproblematically physical; and there is a subject, also physical, who is conscious of the thing.

Is that it? It seems sort of underwhelming. But I fear that is the essence of it. Helpfully, Honderich provides a table of his proposed structure (next page). It seems to me to confirm the suggested interpretation.

So it kind of looks as if Honderich has used a confusingly non-standard definition and ended up with a theoretical position which honestly sheds little light on the real issue; yet these were the very problems he criticised in earlier approaches. I can't deny that I have greatly simplified here and it might be that I missed the key somewhere in one of those many chapters - but frankly I'm not going back to look again.



	OBJECTIVE PHYSICAL WORLD	SUBJECTIVE PHYSICAL WORLDS: Perceptual Consciousness	SUBJECTIVE PHYSICAL REPRESENTATIONS: Cognitive And Affective Consciousness
	ITS PHYSICALITY	***THEIR PHYSICALITY***	***THEIR PHYSICALITY***
1	in the inventory of science	in the inventory of science	in the inventory of science
2	open to the scientific method	open to the scientific method	open to the scientific method
3	in space and time	in space and time	in space and time
4	in particular lawful connections	in particular lawful connections	in particular lawful connections
5	in categorial lawful connections	in categorial lawful connections, including those with the objective physical world and conscious thing	in categorial lawful connections, including those with the objective physical world and conscious thing
6	macroworld perception, microworld deduction	constitutive of macroworld perception	not perceived, but dependent on macroworld perception
7	more than one point of view with macroworld	more than one point of view with with perception	no point of view
8	different from different points of view	different from different points of view	no differences from points of view
9	primary and secondary properties	primary and secondary properties	no primary and secondary properties
	ITS OBJECTIVITY	***THEIR SUBJECTIVITY***	***THEIR SUBJECTIVITY***
10	separate from consciousness	not separate from consciousness	not separate from consciousness
11	public	private	private
12	common access	privileged access	privileged access
13	truth and logic, more subject to	truth and logic, less subject to?	truth and logic, less subject to
14	open to the scientific method	open to the scientific method despite doubt	open to the scientific method despite doubt
15	includes no self or unity or other such inner fact of subjectivity inconsistent with the above properties of the objective physical world	each subjective physical world is an element in an individuality that is a unique and large unity of lawful and conceptual dependencies including much else	each representation is an element in an individuality that is a unique and large unity of lawful and meaningful dependencies including much else
16	hesitation about whether objective physicality includes consciousness	no significant hesitation about taking the above subjective physicality as being that of actual perceptual consciousness	no significant hesitation about taking this subjective physicality as being the nature of actual cognitive and affective consciousness

The Nightmare Scenario

20 December 2014

Microsoft recently announced the first public beta preview for Skype Translate, a service which will provide immediate translation during voice calls. For the time being only Spanish/English is working but we're told that English/German and other languages are on the way. The approach used is complex. Deep Neural Networks apparently play a key role in the speech recognition. While the actual translation ultimately relies on recognising bits of text which resemble those it already knows, the same basic principle applied in existing text translators such as Google Translate, it is also capable of recognising and removing 'disfluencies' - um and ers, rephrasings, and so on, and apparently makes some use of syntactical models, so there is some highly sophisticated processing going on. It seems to do a reasonable job, though as always with this kind of thing a degree of scepticism is appropriate.

Translating actual speech, with all its messy variability is of course an amazing achievement, much more difficult than dealing with text (which itself is no walk in the park); and it's remarkable indeed that it can be done so well without the machine making any serious attempt to deal with the meaning of the words it translates. Perhaps that's a bit too bald: the software does take account of context and as I said it removes some meaningless bits, so arguably it is not ignoring meaning totally. But full-blown intentionality is completely absent.

This fits into a recent pattern in which barriers to AI are falling to approaches which skirt or avoid consciousness as we normally understand it, and all the intractable problems that go with it. It's not exactly the triumph of brute force, but it does owe more to processing power and less to ingenuity than we might have expected. At some point if this continues, we're going to have to take seriously the prospect of our having a machine which mimics human behaviour brilliantly without our ever having solved any of the philosophical problems. Such a robot might run on something like a revival of the frames or scripts of Marvin Minsky or Roger Schank, only this time with a depth and power behind it that would make the early attempts look like working with an abacus. The AI would, at its crudest, simply be recognising situations and looking up a good response, but it would have such a gigantic library of situations and it would be so subtle at customising the details that its behaviour would be entirely indistinguishable from that of ordinary humans. What would we say about such a robot (let's call her Sophia, why not since anthropomorphism seems inevitable). I can see several options.

Option one. Sophia really is conscious, just like us. OK, we don't really understand how we pulled it off, but it's futile to argue about it when her performance provides everything we could possibly demand of consciousness. We don't argue that photographs are not depictions because they're not executed in oil paint, so why would we argue that a consciousness achieved by other means is not the real thing? She achieved consciousness by a different route, and her brain doesn't work like ours - but her mind does. In fact, it turns out we probably work more like her than we thought: all this talk of real intrinsic intentionality and magic

meaningfulness turns out to be a systematic delusion; we're really just running scripts ourselves!

Option two. Sophia is conscious, but not in the way we are. OK, the results are indistinguishable, but we just know that the methods are different, and so the process is not the same. birds and bats both fly, but they don't do it the same way. Sophia probably deserves the same moral rights and duties as us, though we need to be careful about that; but she could very well be a philosophical zombie who has no subjective experience. On the other hand, her mental life might have subjective qualities of its own, very different to ours but incommunicable.

Option three. She's not not conscious; we just know she isn't, because we know how she works and we know that all her responses and behaviour come from simply picking canned sequences out of the cupboard. We're deluding ourselves if we think otherwise. But she is the vivid image of a human being and an incredibly subtle and complex entity: she may not be that different from animals whose behaviour is largely instinctive. We cannot therefore simply treat her as a machine: she probably ought to have some kinds of rights: perhaps special robot rights. Since we can't be absolutely certain that she does not experience real pain and other feelings in some form, and since she resembles us so much, it's right to avoid cruelty both on the grounds of the precautionary principle and so as not to risk debasing our own moral instincts; if we got used to doling out bad treatment to robots who cried out with human voices, we might get used to doing it to flesh and blood people too.

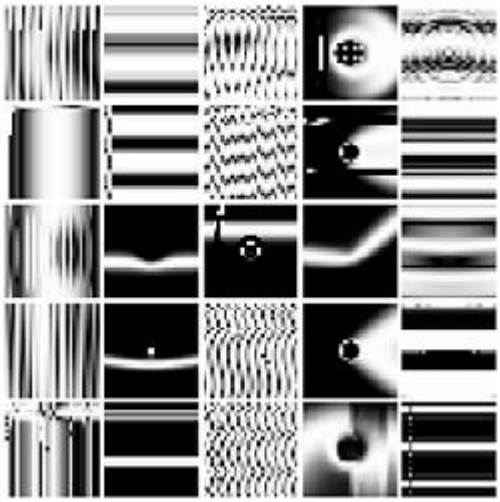
Option four. Sophia's just an entertaining machine, not conscious at all; but that moral stuff is rubbish. It's perfectly OK to treat her like a slave, to turn her off when we want, or put her through terrible 'ordeals' if it helps or amuses us. We know that inside her head the lights are off, no-one home: we might as well worry about dolls. You talk about debasing our moral instincts, but I don't think treating puppets like people is a great way to go, morally. You surely wouldn't switch trolleys to save even ten Sophias if it killed one human being: follow that out to its logical conclusion.

Option five. Sophia is a ghastly parody of human life and should be destroyed immediately. I'm not saying she's actuated by demonic possession (although Satan is pretty resourceful), but she tempts us into diabolical errors about the unique nature of the human spirit.

No doubt there are other options; for me. at any rate, being obliged to choose one is a nightmare scenario.

Illusions For Robots

4 January 2015



Neural networks really seem to be going places recently. Last time I mentioned their use in sophisticated translation software, but they're also steaming ahead with new successes in recognition of visual images. Recently there was a claim from MIT that the latest systems were catching up with primate brains at last. Also from MIT (also via MLU) though, has come an intriguing study into what we could call optical illusions for robots, which cause the systems to make mistakes which are incomprehensible to us primates. The graphics in the grid on the right apparently look like a selection of digits between one and six in the eyes of these recognition systems. Nobody really knows why, because of course neural networks are trained, not programmed, and develop their own inscrutable methods.

How then, if we don't understand, could we ever create such illusions? Optical illusions for human beings exploit known methods of visual analysis used by the brain, but if we don't know what method a neural network is using, we seem to be stymied. What the research team did is use one of their systems in reverse, getting it to create images instead of analysing them. These were then evaluated by a similar system and refined through several iterations until they were accepted with a very high level of certainty.

This seems quite peculiar and the first impression is that it rather seriously undermines our faith in the reliability of neural network systems. However, there's one important caveat to take into account: the networks in question are 'used to' dealing with images in which the crucial part to be identified is small in relation to the whole. They are happy ignoring almost all of the image. So to achieve a fair comparison with human recognition we should perhaps think of the question being not 'do these look like numbers to you?' and more like 'can you find one of the digits from one to six hidden somewhere in this image?'. On that basis the results seem easier to understand.

There still seem to be some interesting implications, though. The first is that, as with language, AI systems are achieving success with methods that do not much resemble those used by the human brain. There's an irony in this happening with neural networks, because in the old dispute between GOFAI and networks it was the network people who were trying to follow a biological design, at least in outline. The opposition wanted to treat cognition as a pure engineering problem; define what we need, identify the best way to deliver it, and don't worry about copying the brain. This is the school of thought that likes to point out that we didn't achieve flight be making machines with flapping, feathery wings. Early network theory, going right back to McCulloch and Pitts, held that we were better off designing something that looked at least broadly like the neurons in the brain. In fact, of course, the resemblance has never been that close, and the focus has generally been more on results than on replicating the structures and systems of biological brains; you could argue that modern neural networks are no more like the brain than fixed-wing aircraft are to birds (or bats). At any rate, the prospect of equalling human performance without doing it the human way raises the same nightmare scenario I was

talking about last time; robots that are not people but get treated as if they were (and perhaps people being treated like machines as a consequence.

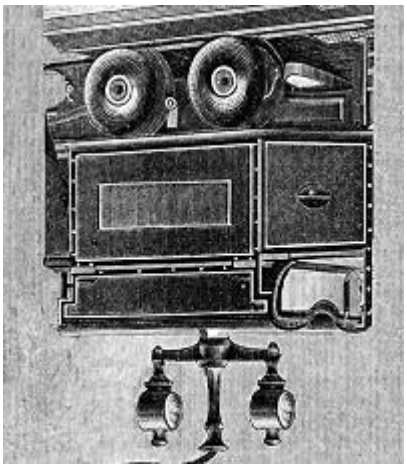
A second issue is whether the deception which these systems fall into points to a general weakness. Could it be that these systems work very well when dealing with 'ordinary' images but continue go wildly off the rails when faced with certain kinds of unusual ones - even when being pout to practical use? It's perhaps not very likely that system is going to run into the kind of truly bizarre image we seem to be dealing with, but a more realistic concern might be the potential scope for sabotage or subversion on the part of some malefactor. One safeguard against this possibility is that the images in question were designed by, as it were, sister systems, ones that worked pretty much the same way and presumably shared the same quirks. Without owning one of these systems yourself it might be difficult to devise illusions that worked - unless perhaps there are general illusions that all network systems are more or less equally likely to be fooled by? That doesn't seem very likely, but it might be an interesting research project. The other safeguard is that these systems are not likely to be used without some additional safeguards, perhaps even more contextual processing of broadly the kind that the human mind obviously brings to the task.

The third question is - what is it like to be an AI deceived by an illusion? There's no reason to think that these machines have subjective experience - unless you're one of those who is prepared to grant a dim glow of awareness to quite simple machines - but what if some cyborg with a human brain, or a future conscious robot, had systems like these as part of its processing apparatus rather than the ones provided by the human brain? It's not implausible that the immense plasticity of the human brain would allow the inputs to be translated into normal visual experience, or something like it. On the whole I think this is the most likely result, although there might be quirks or deficits (or hey, enhancements, why not) in the visual experience. The second possibility is that the experience would be completely weird and inexpressible and although the cyborg/robot would be able to negotiate the world just fine, its experience would be like nothing we've ever had, perhaps like nothing we can imagine.

The third possibility is that it would be like nothing. There would be no experience as such; the data and the knowledge about the surroundings would appear in the cyborg/human's brain but there would be nothing it was like for that to happen. This is the answer qualophile scpetice would expect for a pure robot brain, but the cyborg is more worrying. Human beings are supposed to experience qualia, but when do they arise? Is it only after all the visual processing has been done - when the data ariive in the 'Cartesian Theatre' which Dennett has often told us does not exist? Is it, instead, in the visual processing modules or at the visual processing stage? If so, then we were wrong to doubt that MIT's systems are not having experiences. Perhaps the cyborg gets flawed or partial qualia - but what would that even mean..?

Aliens Are Robots

18 January 2015



Susan Schneider's recent paper argues that when we hear from alien civilisations, it's almost bound to be super intelligent robots getting in touch, rather than little green men. She builds on Nick Bostrom's much-discussed argument that we're all living in a simulation.

Actually, Bostrom's argument is more cautious than that, and more carefully framed. His claim is that at least one of the following propositions is true:

(1) the human species is very likely to go extinct before reaching a "posthuman" stage;

(2) any posthuman civilization is extremely unlikely to run a significant number of simulations of their evolutionary history (or variations thereof);

(3) we are almost certainly living in a computer simulation.

So that if we disbelieve the first two, we must accept the third.

In fact there are plenty of reasons to argue that the first two propositions are true. The first evokes ideas of nuclear catastrophe or an unexpected comet wiping us out in our prime, but equally it could just be that no post human stage is ever reached. We only know about the cultures of our own planet, but two of the longest lived - the Egyptian and the Chinese - were very stable, showing few signs of moving on towards post humanism. They made the odd technological advance, but they also let things slip: no more pyramids after the Old Kingdom; ocean-going junks abandoned before being fully exploited. Really only our current Western culture, stemming from the European Renaissance, has displayed a long run of consistent innovation; it may well be a weird anomaly and its five-hundred year momentum may well be temporary. Maybe our descendants will never go much further than we already have; maybe, thinking of Schneider's case, the stars are basically inhabited by Ancient Egyptians who have been living comfortably for millions of years without ever discovering electricity.

The second proposition requires some very debatable assumptions, notably that consciousness is computable. But the notion of "simulation" also needs examination. Bostrom takes it that a computer simulation of consciousness is likely to be conscious, but I don't think we'd assume a digital simulation of digestion would do actual digesting. The thing about a simulation is that by definition it leaves out certain aspects of the real phenomenon (otherwise it's the phenomenon itself, not a simulation). Computer simulations normally leave out material reality, which could be a problem if we want real consciousness. Maybe it doesn't matter for consciousness; Schneider argues strongly against any kind of biological requirement and it may well be that functional relations will do in the case of consciousness. There's another issue, though; consciousness may be uniquely immune from simulation because of its strange epistemological greediness. What do I mean? Well, for a simulation of digestion we can write a list of all the entities to be dealt with - the foods we expect to enter the gut and their main components. It's not an unmanageable task, and if we like we can leave out some items or some classes of item without thereby invalidating the simulation. Can we write a list of the possible contents of consciousness? No. I can think about any damn thing I like, including

fictional and logically impossible entities. Can we work with a reduced set of mental contents? No; this ability to think about anything is of the essence.

All this gets much worse when Bostrom floats the idea that future ancestor simulations might themselves go on to be post human and run their own nested simulations, and so on. We must remember that he is really talking about simulated worlds, because his simulated ancestors need to have all the right inputs fed to them consistently. A simulated world has to be significantly smaller in information terms than the world that contains it; there isn't going to be room within it to simulate the same world again at the same level of detail. Something has to give.

Without the indefinite nesting, though, there's no good reason to suppose the simulated ancestors will ever outnumber the real people who ever lived in the real world. I suppose Bostrom thinks of his simulated people as taking up negligible space and running at speeds far beyond real life; but when you're simulating everything, that starts to be questionable. The human brain may be the smallest and most economic way of doing what the human brain does.

Schneider argues that, given the same Whiggish optimism about human progress we mentioned earlier, we must assume that in due course fleshy humans will be superseded by faster and more capable silicon beings, either because robots have taken over the reins or because humans have gradually cyborgised themselves to the point where they are essentially super intelligent robots. Since these post human beings will live on for billions of years, it's almost certain that when we make contact with aliens, that will be the kind we meet.

She is, curiously, uncertain about whether these beings will be conscious. She really means that they might be zombies, without phenomenal consciousness. I don't really see how super intelligent beings like that could be without what Ned Block called access consciousness, the kind that allows us to solve problems, make plans, and generally think about stuff; I think Schneider would agree, although she tends to speak as though phenomenal, experiential consciousness was the only kind.

She concludes, reasonably enough, that the alien robots most likely will have full conscious experience. Moreover, because reverse engineering biological brains is probably the quick way to consciousness, she thinks that a particular kind of super intelligent AI is likely to predominate: biologically inspired superintelligent alien (BISA). She argues that although BISAs might in the end be incomprehensible, we can draw some tentative conclusions about BISA minds:

I. Learning about the computational structure of the brain of the species that created the BISA can provide insight into the BISAs thinking patterns.
II. BISAs may have viewpoint invariant representations. (Surely they wouldn't be very bright if they didn't?)
III. BISAs will have language-like mental representations that are recursive and combinatorial. (Ditto.)
IV. BISAs may have one or more global workspaces. (If you believe in global workspace theory, certainly. Why more than one, though - doesn't that defeat the object? Global workspaces are useful because they're global.)
V. A BISA's mental processing can be understood via functional decomposition.

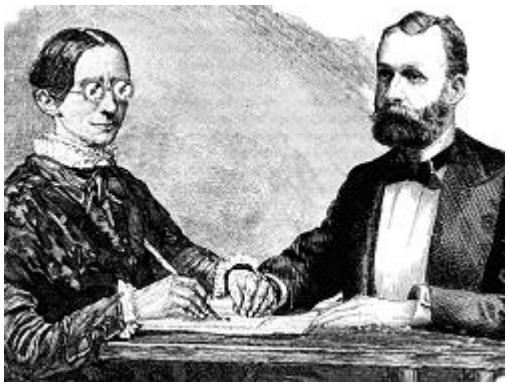
I'll throw in a strange one; I doubt whether BISAs would have identity, at least not the way we do. They would be computational processes in silicon: they could split, duplicate, and merge

without difficulty. They could be copied exactly, so that the question of whether BISA x was the same as BISA y could become meaningless. For them, in fact, communicating and merging would differ only in degree. Something to bear in mind for that first contact, perhaps.

This is interesting stuff, but to me it's slightly surprising to see it going on in philosophy departments; does this represent an unexpected revival of the belief that armchair reasoning can tell us important truths about the world?

Consciousness Speed-Dating

25 January 2015



Why can't we solve the problem of consciousness? That is the question asked by a recent Guardian piece. The account given there is not bad at all; excellent by journalistic standards, although I think it probably overstates the significance of Francis Crick's intervention. His book was well worth reading, but in spite of the title his hypothesis had ceased to be astonishing quite a while before. Surely also a little odd to have Colin McGinn named only as Ted Honderich's adversary when his own Mysterian views are so much more widely cited. Still the piece makes a good point; lots of Davids and not a few Samsons have gone up against this particular Goliath, yet the giant is still on his feet.

Well, if several decades of great minds can't do the job, why not throw a few dozen more at it? The Edge, in its annual question this year, asks its strike force of intellectuals to tackle the question: What do you think about machines that think? This evoked no fewer than 186 responses. Some of the respondents are old hands at the consciousness game, notably Dan Dennett; we must also tip our hat to our friend Arnold Trehub, who briefly denounces the idea that artefactual machines can think. It's certainly true, in my own opinion, that we are nowhere near thinkng machines, and in fact it's not clear that we are getting materially closer: what we have got is splendid machines that clearly don't think at all but are increasingly good at doing tasks we previously believed needed thought. You could argue that eliminating the need for thought was Babbage's project right from the beginning, and we know that Turing discarded the question 'Can machines think?' as not worthy of an answer.

186 answers is of course, at least 185 more than we really wanted, and those are not good odds of getting even a congenial analysis. In fact, the rapid succession of views, some well-informed, others perhaps shooting from the hip to a degree, is rather exhausting: the effect is like a dreadfully prolonged session of speed dating: like my theory? No? Well don't worry, there are 180 more on the way immediately. It is sort of fun to surf the wave of punditry, but I'd be surprised to hear that many people were still with the programme when it got to view number 186 (which, despairingly or perhaps refreshingly, is a picture).

Honestly. though, why can't we solve the problem of consciousness? Could it be that there is something fundamentally wrong? Colin McGinn, of course, argues that we can never understand consciousness because of cognitive closure; there's no real mystery about it, but our mental toolset just doesn't allow us to get to the answer. McGinn makes a good case, but I think that human cognition is not formal enough to be affected by a closure of this kind; and if it were, I think we should most likely remain blissfully unaware of it: if we were unable to understand consciousness, we shouldn't see any problem with it either.

Perhaps, though, the whole idea of consciousness as conceived in contemporary Western thought is just wrong? It does seem to be the case that non-European schools of philosophy construe the world in ways that mean a problem of consciousness never really arises. For that matter, the ancient Greeks and Romans did not really see the problem the way we do: although ancient philosophers discussed the soul and personal identity, they didn't really worry about

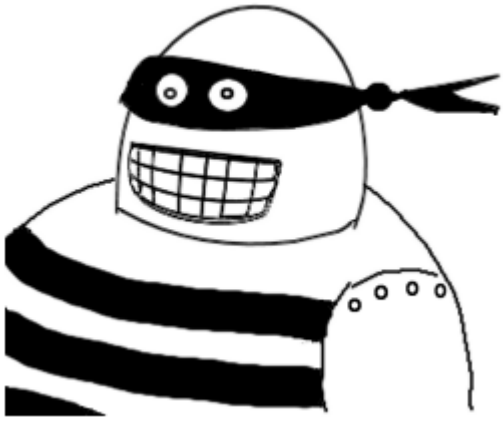
consciousness. Commonly people blame Western dualism for drawing too sharp a division between the world of the mind and the world of material objects: and the finger is usually pointed at Descartes in particular. Perhaps if we stopped thinking about a physical world and a non-physical mind the alleged problem would simply evaporate. If we thought of a world constituted by pure experience, not differentiated into two worlds, everything would seem perfectly natural?

Perhaps, but it's not a trick I can pull off myself. I'm sure it's true our thinking on this has changed over the years, and that the advent of computers, for example, meant that consciousness, and phenomenal consciousness in particular, became more salient than before. Consciousness provided the extra thing computers hadn't got, answering our intuitive needs and itself being somewhat reshaped to fill the role. William James, as we know, thought the idea was already on the way out in 1904: "A mere echo, the faint rumour left behind by the disappearing 'soul' upon the air of philosophy"; but over a hundred years later it still stands as one of the great enigmas.

Still, maybe if we send in another 200 intellectuals…?

Crimbots

2 February 2015



Some serious moral dialogue about robots recently. Eric Schwitzgebel put forward the idea that we might have special duties in respect of robots, on the model of the duties a parent owes to children, an idea embodied in a story he wrote with Scott Bakker. He followed up with two arguments for robot rights; first, the claim that there is no relevant difference between humans and AIs, second, a Bostromic argument that we could all be sims, and if we are, then again, we're not different from AIs.

Scott has followed up with a characteristically subtle and bleak case for the idea that we'll be unable to cope with the whole issue anyway. Our cognitive capacities, designed for shallow information environments, are not even up to understanding ourselves properly; the advent of a whole host of new styles of cognition will radically overwhelm them. It might well be that the revelation of how threadbare our own cognition really is will be a kind of poison pill for philosophy (a well-deserved one on this account, I suppose).

I think it's a slight mistake to suppose that morality confers a special grade of duty in respect of children. It's more that parents want to favour their children, and our moral codes are constructed to accommodate that. It's true society allocates responsibility for children to their parents, but that's essentially a pragmatic matter rather than a directly moral one. In wartime Britain the state was happy to make random strangers responsible for evacuees, while those who put the interests of society above their own offspring, like Brutus (the original one, not the Caesar stabber) have sometimes been celebrated for it.

What I want to do though, is take up the challenge of showing why robots are indeed relevantly different to human beings, and not moral agents. I'm addressing only one kind of robot, the kind whose mind is provided by the running of a program on a digital computer (I know, John Searle would be turning in his grave if he wasn't still alive, but bear with me). I will offer two related points, and the first is that such robots suffer grave problems over identity. They don't really have personal identity, and without that they can't be moral agents.

Suppose Crimbot 1 has done a bad thing; we power him down, download his current state, wipe the memory in his original head, and upload him into a fresh robot body of identical design.

"Oops, I confess!" he says. Do we hold him responsible; do we punish him? Surely the transfer to a new body makes no difference? It must be the program state that carries the responsibility; we surely wouldn't punish the body that committed the crime. It's now running the Saintbot program, which never did anything wrong.

But then neither did the copy of Crimbot 1 software which is now running in a different body - because it's a copy, not the original. We could upload as many copies of that as we wanted; would they all deserve punishment for something only one robot actually did?

Maybe we would fall back on the idea that for moral responsibility it has to be the same copy in the same body? By downloading and wiping we destroyed the person who was guilty and merely created an innocent copy? Crimbot 1 in the new body smirks at that idea.

Suppose we had uploaded the copy back into the same body? Crimbot 1 is now identical, program and body, the same as if we had merely switched him off for a minute. Does the brief interval when his data registers had different values make such a moral difference? What if he downloaded himself to an internal store, so that those values were always kept within the original body? What if he does that routinely every three seconds? Does that mean he is no longer responsible for anything, (unless we catch him really quickly) while a version that doesn't do the regular transfer of values can be punished?

We could have Crimbot 2 and Crimbot 3; 2 downloads himself to internal data storage every second and the immediately uploads himself again. 3 merely pauses every second for the length of time that operation takes. Their behaviour is identical, the reasons for it are identical; how can we say that 2 is innocent while 3 is guilty?

But then, as the second point, surely none of them is guilty of anything? Whatever may be true of human beings, we know for sure that Crimbot 1 had no choice over what to do; his behaviour was absolutely determined by the program. If we copy him into another body, and set him uip wioth the same circumstances, he'll do the same things. We might as well punish him in advance; all copies of the Crimbot program deserve punishment because the only thing that prevented them from committing the crime would be circumstances.

Now, we might accept all that and suggest that the same problems apply to human beings. If you downloaded and uploaded us, you could create the same issues; if we knew enough about ourselves our behaviour would be fully predictable too!

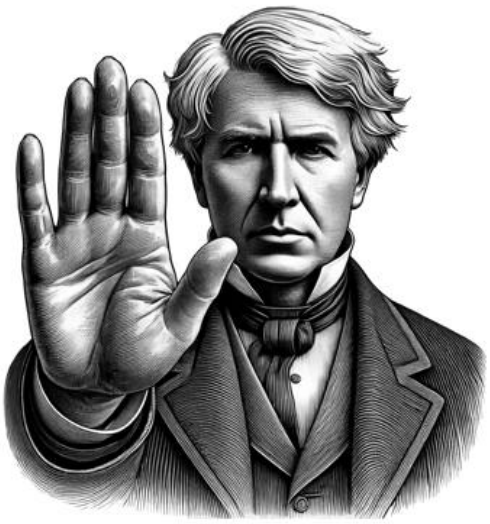
The difference is that in Crimbot the distinction between program and body is clear because he is an artefact, and he has been designed to work in certain ways. We were not designed, and we do not come in the form of a neat layer of software which can be peeled off the hardware. The human brain is unbelievably detailed, and no part of it is irrelevant. The position of a single molecule in a neuron, or even in the supporting astrocytes, may make the difference between firing and not firing, and one neuron firing can be decisive in our behaviour. Whereas Crimbot's behaviour comes from a limited set of carefully designed functional properties, ours comes from the minute specifics of who we are. Crimbot embodies an abstraction, he's actually designed to conform as closely as possible to design and program specs; we're unresolvably particular and specific.

Couldn't that, or something like that, be the relevant difference?

Minimising Free Energy

8 February 2015

A real wealth of papers at the OpenMind site, presided over by Thomas Metzinger, including new stuff from Dan Dennett, Ned Block, Paul Churchland, Alva Noë, Andy Clark and many others. Call me perverse, but the one that attracted my attention first is the paper The Neural Organ Explains the Mind by Jakob Hohwy. This expounds the free energy theory put forward by Karl Friston.



Edison Says No

The hypothesis here is that we should view the brain as an organ of the body in the same way as we regard the heart or the liver. Those other organs have a distinctive function - in the case of the heart, it pumps blood - and what we need to do is recognise what the brain does. The suggestion is that it minimises free energy; that honestly doesn't mean much to me, but apparently another way of putting it is to say that the brain's function is to keep the organism within a limited set of states. If the organism is a fish, the brain aims to keep it in the right kind of water, keeps it fed and with adequate oxygen, and so on.

It's always good to get back to some commonsensical. pragmatic view, and the paper shows that this is a fertile and flexible hypothesis, which yields explanations for various kinds of behaviour. There seem to me to be three prima facie objections. First, this isn't really a function akin to the heart pumping blood; at best it's a high level meta-function. The heart does blood pumping, the lungs do respiration, the gut does digestion; and the brain apparently keeps the organism in conditions where blood can go on being pumped, there is still breathable air available, food to be digested, and so on. In fact it oversees every other function and in completely novel circumstances it suddenly acquires new functions: if we go hang-gliding, the brain learns to keep us flying straight and level, not something it ever had to do in the earlier history of the human race. Now of course, if we confront the gut with a substance it never experienced before, it will probably deal with it one way or another; but it will only deploy the chemical functions it always had; it won't learn new ones. There's a protean quality about the brain that eludes simple comparisons with other organs.

A second problem is that the hypothesis suggests the brain is all about keeping the organism in states where it is comfortable, whereas the human brain at least seems to be able to take into account future contingencies and make long-term plans which enable us to abandon the ideal environment of our beds each morning and go out into the cold and rain. There is a theoretical answer to this problem which seems to involve us being able to perceive things across space and time; probably right, but that seems like a whole new function rather then something that drops out naturally from minimising free energy; I may not have understood this bit correctly. It seems that when we move our hand, it may happen because we have, in contradiction of the evidence, adopted the belief that our hand is already moving; this belief serves to minimise free energy and our belief that the hand is moving causes the actual movement we believe in.

Third and worse, the brain often seems to impel us to do things that are risky, uncomfortable, and damaging, and not necessarily in pursuit of keeping our states in line with comfort, even in the long term. Why do we go hang-gliding, why do we take drugs, climb mountains, enter a monastery or convent? I know there are plenty of answers in terms of self-interest, but it's much less clear to me that there are answers in terms of minimising free energy.

That's all very negative, but actually the whole idea overall strikes me as at least a novel and interesting perspective. Hohwy draws some parallels with the theory of evolution; like Darwin's idea, this is a very general theory with alarmingly large claims, and critics may well say that it's either over-ambitious or that in the end it explains too much too easily; that it is, ultimately, unfalsifiable.

I wouldn't go that far; it seems to me that there are a lot of potential issues, but that the theory is very adaptable and productive in potentially useful ways. It might well be a valuable perspective. I'm less sure that it answers the questions we're really bothered about. Take the analogy of the gut (as the theory encourages us to do). What is the gut's function? Actually, we could define it several ways (it deals with food, it makes poop, it helps store energy). One might be that the gut keeps the bloodstream in good condition so far as nutrients are concerned, just as the lungs keep it in good condition in respect of oxygenation. But the gut also, as part of that, does digestion, a complex and fascinating subject which is well worth study in itself. Now it might be that the brain does indeed minimise free energy, and that might be a legitimate field of study; but perhaps in doing so it also supports consciousness, a separate issue which like digestion is well worthy of study in itself.

We might not be looking at final answers, then - to be fair, we've only scratched the surface of what seems to be a remarkably fecund hypothesis - but even if we're not, a strange new idea has got to be welcome.

The Stoppard Problem

15 February 2015

It was exciting to hear that Tom Stoppard's new play was going to be called The Hard Problem, although until it opened recently details were scarce. In the event the reviews have not been very good. It could easily have been that the pieces in the mainstream newspapers missed the point in some way; unfortunately, Vaughan Bell of Mind Hacks didn't like the way the intellectual issues were handled either (though he had an entertaining evening); and he's a very sensible and well-informed commentator on consciousness and the mind. So, a disappointing late entry in a distinguished playwright's record?

I haven't seen it yet, but I've read the script, which in some ways is better for our current purposes. No-one, of course, supposed that Stoppard was going to present a live solution to the Hard Problem: but in the event the play is barely about that problem at all. The Problem's chief role is to help Hilary, our heroine, get a job at the Krohl Institute for Brain Science, an organisation set up by the wealthy financier Jerry Krohl. Most of the Krohl's work is on 'hard' neuroscience and reductive, materialist projects, but Leo, the head of the department Hilary joins, happens to think the Hard Problem is central. Merely mentioning it is enough to clinch the job, and that's about it; the chief concern of the rest of the research we're told about is altruism, and the Prisoner's Dilemma.

The strange thing is that within philosophy the Hard Problem must be the most fictionalised issue ever. The wealth of thought experiments, elaborate examples and complicated counterfactuals provides enough stories to furnish the complete folklore of a small country. Mary the colour scientist, the zombies, the bats, Twin Earth, chip-head, the qualia that dance and the qualia that fade like Tolkienish elves; as an author you'd want to make something out of all that, wouldn't you? Or perhaps that assumption just helps explain why I'm not a successful playwright. Of course, you'd only use that stuff if you really wanted to write about the Hard Problem, and Stoppard, it seems, doesn't really. Perhaps he should just have picked a different title; Every Good Girl Deserves to Know What Became of Her Kid?

Hilary, in fact, had a daughter as a teenager who she gave up for adoption, and who she has worried about ever since. She believes in God because she needs someone effective to pray to about it; and presumably she believes in altruism so someone can be altruistic towards her daughter; though if the sceptic's arguments are sound, self-interest would work, too.

The debate about altruism is one of those too-well-trodden paths in philosophy; more or less anything you say feels as if it has been in a thousand mouths already. I often feel there's an element of misunderstanding between those who defend the concept of altruism and those who would reduce it to selfish genery. Yes, the way people behave tends to be consistent with their own survival and reproduction; but that hardly exhausts the topic; we want to know how the actual reasons, emotions, and social conventions work. It's sort of as though I remarked on how extraordinary it is that a forest pumps so much water way above the ground.¶

> *"There's no pump, Peter," says BitBucket; "that's kind of a naive view. See, the tree needs the water in its leaves to survive, so it has evolved as a water-having organism. There are no little hamadryads planning it all out and working tiny pumps. No water magic."*
>
> *"But there's like, capillarity, or something, isn't there? Um, osmosis? Xylem and phloem? Turgid vacuoles?"*
>
> *"Sure, but those things are completely explained by the evolutionary imperatives. Saying there are vacuoles doesn't tell us why there are vacuoles or why they are what they really are."*

"I don't think osmosis is ***completely*** *explained by evolution. And surely the biological pumping system is, you know, worth discussing in itself?"*

"There's no pump, Peter!"

Stoppard seems to want to say that greedy reductionism throws out the baby with the bath water. Hilary's critique of the Prisoner's Dilemma is that it lacks all context, all the human background that actually informs our behaviour; what's the relationship of the two prisoners? When the plot puts her into an analogous dilemma, she sacrifices her own interests and career, and is forced to suffer the humiliation of being left with nothing to study but philosophy. In parallel the financial world that pays for the Krohl is going through convulsions because it relied on computational models which were also too reductionist; it turns out that the market thinks and feels and reacts in ways that aren't determined by rational game theory.

That point is possibly a little undercut by the fact that a reductionist actually foresaw the crash. Amal, who lost out by not rating the Hard Problem high enough, nevertheless manages to fathom the market problem ahead of time…¶

The market is acting stupid, and the models are out of whack because we don't know how to build a stupid computer.

But perhaps we are to suppose that he's learnt his lesson and is ready to talk turgid vacuoles with us sloppy thinkers.

I certainly plan to go and see a performance, and if new light dawns as a result, I'll let you know.

Time Travel Consciousness

22 February 2015

Can you change your mind after the deed is done? Ezequiel Di Paolo thinks you can, sometimes. More specifically, he believes that acts can become intentional after they have already been performed. His theory, which seems to imply a kind of time travel, is set out in a paper in the latest JCS.

I think the normal view would be that for an act to be intentional, it must have been caused by a conscious decision on your part. Since causes come before effects, the conscious decision must have happened beforehand, and any thoughts you may have afterwards are irrelevant. There is a blurry borderline over what is conscious, of course; if you were confused or inattentive, if you were 'on autopilot' or you were following a hunch or a whim it may not be completely clear how consciously your action was considered.

There can, moreover, be what Di Paolo calls an epistemic change. In such a case the action was always intentional in fact, but you only realise that it was when you think about your own motives more carefully after the event. Perhaps you act in the heat of the moment without reflection; but when you think about it you realise that in fact what you did was in line with your plans and actually caused by them. Although this kind of thing raises a few issues, it is not deeply problematic in the same way as a real change. Di Paolo calls the real change an ontological one; here you definitely did not intend the action beforehand, but it becomes intentional retrospectively.

That seems disastrous on the face of it. If the intentionality of an act can change once, it can presumably change again, so it seems all intentions must become provisional and unreliable; the whole concept of responsibility looks in danger of being undermined. Luckily, Di Paolo believes that changes can only occur in very particular circumstances, and in such a way that only one revision can occur.

His view founds intentions in enactment rather than in linear causation; he has them arising in social interaction. The theory draws on Husserl and Heidegger, but probably the easiest way to get a sense of it is to consider the examples presented by Di Paolo. The first is from De Jaegher and centres, in fittingly continental style, around a cheese board.

De Jaegher is slicing himself a corner of Camembert and notices that his companion is watching in a way which suggests that he too, would like to eat cheese. DJ cuts him a slice and hands it over.

> "I could see you wanted some cheese," he remarks.
>
> "Funny thing, that," he replies, "actually, I wasn't wanting cheese until you handed it to me; at that moment the desire crystallised and I now found I had been wanting cheese."

In a corner of the room, Alice is tired of the party to do; the people are boring and the magnificent cheese board is being monopolised by philosophers enacting around it. She looks across at her husband and happens to scratch her wrist. He comes over.

> "Saw you point at your watch," he says, "yeah, we probably should go now. We've got the Stompers' do to go to."

Alice now realises that although she didn't mean to point to her watch originally, she now feels the earlier intention is in place after all - she did mean to suggest they went.

At the Stompers' there is dancing; the tango! Alice and Bill are really good, and as they dance Bill finds that his moves are being read and interpreted by Alice superbly; she conforms and shapes to match him before he has actually decided what to do; yet she has read him correctly and he realises that after the fact his intentions really were the ones she divined. (I sort of melded the examples.)

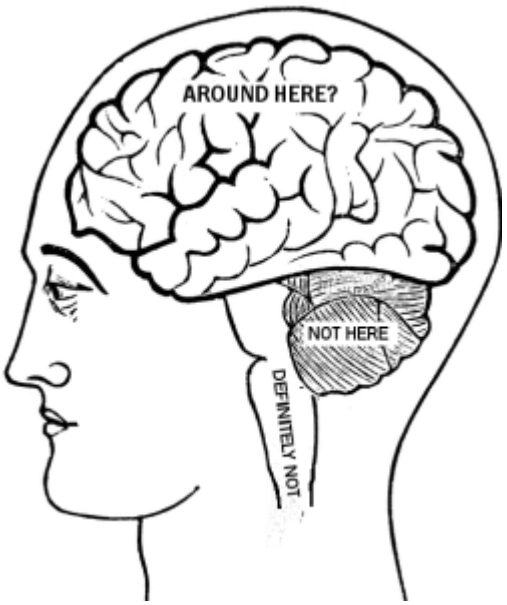
You see how it works? No, it doesn't really convince me either. It is a viable way of looking at things, but it doesn't compel us to agree that there was a real change of earlier intention. Around the cheese board there may always have been prior hunger, but I don't see why we'd say the intention existed before accepting the cheese.

It is true, of course, that human beings are very inclined to confabulate, to make up stories about themselves that make their behaviour make sense, even if that involves some retrospective monkeying with the facts. It might well be that social pressure is a particularly potent source of this kind of thing; we adjust our motivations to fit with what the people around us would like to hear. In a loose sense, perhaps we could even say that our public motives have a social existence apart from the private ones lodged in the recesses of our minds; and perhaps those social ones can be adjusted retrospectively because, to put it bluntly, they are really a species of fiction.

Otherwise I don't see how we can get more than an epistemic change. I've just realised that I really kind of feel like some cheese.

Still Waiting For The NCCs

1 March 2015



An editorial piece in *Frontiers* notes that we still haven't nailed down the neural correlates of consciousness (NCCs). It's part of a Research Topic collection on the subject, and it mentions three candidates featured in the papers which have been well-favoured but now - arguably at any rate - seem to have been found wanting. This old but still useful paper by David Chalmers lists several more of the old contenders. Though naturally a little downbeat, the editorial piece addresses some of the problems and recommends a fresh assault. However, if we haven't succeeded after twenty-five or thirty years of trying, perhaps common sense suggests that there might be something fundamentally wrong with the project?

There must be neural correlates of consciousness, though, mustn't there? Unless we're dualists, and perhaps even if we are, it seems hard to imagine that mental events are not matched by events in the brain. We have by now a wealth of evidence that stimulating parts of the brain can generate conscious experiences artificially, and we've always known that damage to the brain damages the mind; sometimes in exquisitely particular ways. So what could be wrong with the basic premise that there are neural correlates of consciousness?

First, consciousness could itself be a mixed bag of different things, not one consistent phenomenon. Conscious states, after all, include such things as being visually aware of a red light; rehearsing a speech mentally; meditating; and waiting for the starting pistol. These things are different in themselves and it's not particularly likely that their neuronal counterparts will resemble each other.

Then it could be realised in multiple ways. Even if we confine ourselves to one kind of consciousness, there's no guarantee that the brain always does it the same way. If we assume for the sake of argument that consciousness arises from a neuronal function, then perhaps several different processes will do, just as a bucket, a hose, a fountain and a sewer all serve the function of moving water.

Third, it could well be that consciousness arises, not from any property of the neurons doing the thinking, but from the context they do it in. If the higher order theorists were right, to take one example, for a set of neurons to be conscious would require that another set of neurons was directed at them - so that there was a thought about the thought But whether another set of neurons is executing a function about our first set of neurons is not an observable property of the first set of neurons. As another example it might be that theories of embodiment are true in a strong sense, implying that consciousness depends on context external to the brain altogether.

Fourth, consciousness might depend on finely detailed properties that require very complex decoding. Suppose we have a library and we want to find out which books in it mention libraries;

we have to read them to find out. In a somewhat similar way we might have to read the neurons in our brain in detail to find out whether they were supporting consciousness.

Quite apart from these problems of principle, of course, we might reasonably have some reservations about the technology. Even the best scanners have their limitations, typically showing us proxies for the general level of activity in a broad area rather than pinpointing the activity of particular neurons; and it isn't feasible or ethical to fill a subject's brain with electrodes. With the equipment we had twenty-five years ago, it was staggeringly ambitious to think we could crack the problem, but even now we might not really be ready.

All that suggests that the whole idea of Neural Correlates of Consciousness is framed in a way which makes it unpromising or completely misconceived. And yet… understanding consciousness, for most people, is really a matter of building a bridge between the physical and the mental; even if we're not out to reduce the mental to the physical, we want to see, as it were, diplomatic relations established between the two. How could that bridge ever be built without some work on the physical side, and how could that work not be, in part at least, about tracking neuronal activity? If we're not going to succumb to mystery or magic, we just have to keep looking, don't we?

I think there are probably two morals to be drawn. The first is that while we have to keep looking for neural correlates of consciousness in some sense (even if we don't describe the porject that way), it was probably always a little naive to look for the correlates, the single simple things that would infallibly diagnose the presence of consciousness. It was always a bit unlikely, at any rate, that something as simple as oscillating together at 40 Hertz just was consciousness; surely it's was always going to be a lot more complicated than that?

Second, we probably do need a bit more of a theory, or at least a hypothesis. There's no need to be unduly narrow-minded about our scientific method; sometimes even random exploration can lead to significant insights just as well as carefully constructed testing of well-defined hypotheses. But the neuronal activity of the brain is often, and quite rightly, described as the most complex phenomenon in the known universe. Without any theoretical insight into how we think neuronal activity might be giving rise to consciousness, we really don't have much chance of seeing what we're after unless it just happens by great good fortune to be blindingly obvious. Just having a bit of a look to see if we can spot things that reliably occur when consciousness is present is probably underestimating the task. Indeed, that is sort of the theme of the collection; Beyond the Simple Contrastive Approach. To put it crudely, if you're looking for something, it helps to have an idea of what the thing you're looking for looks like.

In another 25 or 30 years, will we still be looking? Or will we have given up in despair? Nil Desperandum!

Hard Not New

9 March 2015

The Hard Problem may indeed be hard, but it ain't new:¶

> *Twenty years ago, however, an instant myth was born: a myth about a dramatic resurgence of interest in the topic of consciousness in philosophy, in the mid-1990s, after long neglect*.

So says Galen Strawson in the TLS: philosophers have been talking about consciousness for centuries. Most of what he says, including his main specific point, is true, and the potted history of the subject he includes is good, picking up many interesting and sensible older views that are normally overlooked (most of them overlooked by me, to be honest). If you took all the papers he mentioned and published them together, I think you'd probably have a pretty good book about consciousness. But he fails to consider two very significant factors and rather over-emphasises the continuity of discussion in philosophy and psychology, leaving a misleading impression.

First, yes, it's absolutely a myth that consciousness came back to the fore in philosophy only in the mid-1990s, and that Francis Crick's book The Astonishing Hypothesis was important in bringing that about. The allegedly astonishing hypothesis, identifying mind and brain, had indeed been a staple of philosophical discussion for centuries. We can also agree that consciousness really did go out of fashion at one stage: Strawson grants that the behaviourists excluded consciousness from consideration, and that as a result there really was an era when it went through a kind of eclipse.

He rather underplays that, though, in two ways. First, he describes it as merely a methodological issue. It's true that the original behaviourists stopped just short of denying the reality of consciousness, but they didn't merely say 'let's approach consciousness via a study of measurable behaviour', they excluded all reference to consciousness from psychology, an exclusion that was meant to be permanent. Second, the leading behaviourists were just the banner bearers for a much wider climate of opinion that clearly regarded consciousness as bunk, not just a non-ideal methodological approach. Interestingly, it looks to me as if Alan Turing was pretty much of this mind. Strawson says:¶

> *But when Turing suggests a test for when it would be permissible to describe machines as thinking, he explicitly puts aside the question of consciousness*.

Actually Turing barely mentions consciousness; what he says is...¶

> *The original question, "Can machines think?" I believe to be too meaningless to deserve discussion*.

The question of consciousness must be at least equally meaningless in his eyes. Turing here sounds very like a behaviourist to me.

What he does represent is the appearance of an entirely new element in the discussion. Strawson represents the history as a kind of debate within psychology and philosophy: it may have been like that at one stage: a relatively civilised dialogue between the elder subject and its

offspring. They'd had a bit of a bust-up when psychology ran away from home to become a science, but they were broadly friends now, recognising each other's prerogatives, and with a lot of common heritage. But in 1950, with Turing's paper, a new loutish figure elbowed its way onto the table: no roots in the classics, no long academic heritage, not even really a science: Artificial Intelligence. But the new arrival seized the older disciplines by the throat and shook them until their teeth rattled, threatening to take the whole topic away from them wholesale. This seminal, transformational development doesn't feature in Strawson's account at all. His version makes it seem as if the bitchy tea-party of philosophy continued undisturbed, while in fact after the rough intervention of AI, psychology's muscular cousin neurology pitched in and something like a saloon bar brawl ensued, with lots of disciplines throwing in the odd punch and even the novelists and playwrights hitching up their skirts from time to time and breaking a bottle over somebody's head.

The other large factor he doesn't discuss is the religious doctrine of the soul. For most of the centuries of discussion he rightly identifies, one's permitted views about the mind and identity were set out in clear terms by authorities who in the last resort would burn you alive. That has an effect. Descartes is often criticised for being a dualist; we have no particular reason to think he wasn't sincere, but we ought to recognise that being anything else could have got him arrested. Strawson notes that Hobbes got away with being a materialist and Hume with saying things that strongly suggested atheism; but they were exceptions, both in the more tolerant (or at any rate more disorderly) religious environment of Britain.

So although Strawson's specific point is right, there really was a substantial sea change: earlier and more complex, but no less worthy of attention.

In those long centuries of philosophy, consciousness may have got the occasional mention, but the discussion was essentially about thought, or the mind. When Locke mentioned the inverted spectrum argument, he treated it only as a secondary issue, and the essence of his point was that the puzzle which was to become the Hard Problem was nugatory, of no interest or importance in itself.

Consciousness per se took centre stage only when religious influence waned and science moved in. For the structuralists like Wundt it was central, but the collapse of the structuralist project led directly to the long night of behaviourism we have already mentioned. Consciousness came back into the centre gradually during the second half of the twentieth century, but this time instead of being the main object of attention it was pressed into service as the last defence against AI; the final thing that computers couldn't do. Whereas Wundt had stressed the scientific measurement of consciousness its unmeasurability was now the very thing that made it interesting. This meant a rather different way of looking at it, and the gradual emergence of qualia for the first time as the real issue. Strawson is quite right of course that this didn't happen in the mid-nineties; rather, David Chalmers' formulation cemented and clarified a new outlook which had already been growing in influence for several decades.

So although the Hard Problem isn't new, it did become radically more important and central during the latter part of the last century; and as yet the sherriff still ain't showed up.

Why No AGI?

15 March 2015

An interesting piece in Aeon by David Deutsch. There was a shorter version in the Guardian, but it just goes to show how even reasonably intelligent editing can mess up a piece. There were several bits in the Guardian version where I was thinking to myself: ooh, he's missed the point a bit there, he doesn't really get that: but on reading the full version I found those very points were ones he actually understood very well. In fact he talks a lot of sense and has some real insights.

Not that everything is perfect. Deutsch quite reasonably says that AGI, artificial general intelligence, machines that think like people, must surely be possible. We could establish that by merely pointing out that if the brain does it, then it seems natural that a machine must be able to do it: but Deutsch invokes the universality of computation, something he says he proved in the 1980s. I can't claim to understand all this in great detail, but I think what he proved was the universality in principle of quantum computation: but the notion of computation used was avowedly broader than Turing computation. So it's odd that he goes on to credit Babbage with discovering the idea, as a conjecture, and Turing with fully understanding it. He says of Turing:¶

> *He concluded that a computer program whose repertoire included all the distinctive attributes of the human brain — feelings, free will, consciousness and all — could be written*.

That seems too sweeping to me: it's not unlikely that Turing did believe those things, but they go far beyond his rather cautious published claims, something we were sort of talking about last time.

I'm not sure I fully grasp what people mean when they talk about the universality of computation. It seems to be that they mean any given physical state of affairs can be adequately reproduced, or at any rate emulated to any required degree of fidelity, by computational processes. This is probably true: what it perhaps overlooks is that for many commonplace entities there is no satisfactory physical description. I'm not talking about esoteric items here: think of a vehicle, or to be Wittgensteinian, a game. Being able to specify things in fine detail, down to the last atom, is simply no use in either case. There's no set of descriptions of atom placement that defines all possible vehicles (virtually anything can be a vehicle) and certainly none for all possible games, which given the fogginess of the idea, could easily correspond with any physical state of affairs. These items are defined on a different level of description, in particular one where purposes and meanings exist and are relevant. So unless I've misunderstood, the claimed universality is not as universal as we might have thought.

However, Deutsch goes on to suggest, and quite rightly, I think, that what programmed AIs currently lack is a capacity for creative thought. Endowing them with this, he thinks, will require a philosophical breakthrough. At the moment he believes we still tend to believe that new insights come from induction; whereas ever since Hume there has been a problem over induction, and no-one knows how to write an algorithm which can produce genuine and reliable new inductions.

Deutsch unexpectedly believes that Popperian epistemology has the solution but has been overlooked. Popper, of course, took the view that scientific method was not about proving a theory but about failing to disprove one: so long as your hypotheses withstood all attempts to prove them false (and so long as they were not cast in cheating ways that made them unfalsifiable) you were entitled to hang on to them.

Maybe this helps to defer the reckoning so far as induction is concerned: it sort of kicks the can down the road indefinitely. The problem, I think, is that the Popperian still has to be able to identify which hypotheses to adopt in the first place; there's a very large if not infinite choice of possible ones for any given set of circumstances.

I think the answer is recognition: I think recognition is the basic faculty underlying nearly all of human thought. We just recognise that certain inductions, and certain events are that might be cases of cause and effect are sound examples: and our creative thought is very largely powered by recognising aspects of the world we hadn't spotted before.

The snag is, in my view, that recognition is unformalisable and anomic - lacking in rules. I have a kind of proof of this. In order to apply rules, we have to be able to identify the entities to which the rules should be applied. This identification is a matter of recognising the entities. But recognition cannot itself be based on rules, because that would then require us to identify the entities to which those rules applied - and we'd be caught in a in a vicious circle.

It seems to follow that if no rules can be given for recognition, no algorithm can be constructed either, and so one of the basic elements of thought is just not susceptible to computation. Whether quantum computation is better at this sort of thing than Turing computation is a question I'm not competent to judge, but I'd be surprised if the idea of rule-free algorithms could be shown to make sense for any conception of computation.

So that might be why AGI has not come along very quickly. Deutsch may be right that we need a philosophical breakthrough, although one has to have doubts about whether the philosophers look likely to supply it: perhaps it might be one of those things where the practicalities come first and then the high theory is gradually constructed after the fact. At any rate Deutsch's piece is a very interesting one, and I think many of his points are good. Perhaps if there were a book-length version I'd find that I actually agree with him completely.

A Book What I Wrote

21 March 2015

It had to happen eventually. I decided it was time I nailed my colours to the mast and said what I actually think about consciousness in book form: and here it is, available on Amazon. *The Shadow of Consciousness (A Little Less Wrong)* has two unusual merits for a book about consciousness: it does not pretend to give the absolute final answer about everything; and more remarkable than that, it features no pictures at all of glowing brains.

Actually, it falls into three parts (only metaphorically – this is a sturdy paperback product or a sound Kindle ebook, depending on your choice). The first is a quick and idiosyncratic review of the history of the subject. I begin with consciousness seen as the property of things that move without being pushed (an elegant definition and by no means the worst) and well, after that it gets a bit more complicated.

The underlying theme here is how the question itself has changed over time, and crucially become less a matter of intellectual justifications and more a matter of practical blueprints for robots. The robots are generally misconceived, and may never really work – but the change of perspective has opened up the issues in ways that may be really helpful.

The second part describes and solves the Easy Problem. No, come on. What it really does is look at the unforeseen obstacles that have blocked the path to AI and to a proper understanding of consciousness. I suggest that a series of different, difficult problems are all in the end members of a group, all of which arise out of the inexhaustibility of real-world situations. The hard core of this group is the classical non-computability established for certain problems by Turing, but the Frame Problem, Quine's indeterminacy of translation, the problem of relevance, and even Hume's issues with induction, all turn out to be about the inexhaustible complexity of the real world.

I suggest that the brain uses the pre-formal, anomic (rule-free) faculty of recognition to deal with these problems, and that that in turn is founded on two special tools; a pointing ability which we can relate to HP Grice's concept of natural meaning, and a doubly ambiguous approach to pattern matching which is highlighted by Edelman's analogy with the immune system.

The third part of the book tackles the Hard Problem. It flails around for quite a while, failing to make much sense of qualia, and finally suggests that in fact there is only one quale; that is, that the special vividness and particularity of real experience which is attributed to qualia is in fact simply down to the haecceity – the 'thisness' of real experience. In the classic qualia arguments, I suggest, we miss this partly because we fail to draw the correct distinction between existence and subsistence (honestly the point is not as esoteric as it sounds).

Along the way I draw some conclusions about causality and induction and how our clerkish academic outlook may have led us astray now and then.

Not many theories have rated more than a couple of posts on Conscious Entities, but I must say I've rather impressed myself with my own perspicacity, so I'm going to post separately about four of the key ideas in the book, alternating with posts about other stuff. The four ideas are inexhaustibility, pointing, haecceity and reality. Then I promise we can go back to normal.

I'll close by quoting from the acknowledgements...

> *... it would be poor-spirited of me indeed not to tip my hat to the regulars at Conscious Entities, my blog, who encouraged and puzzled me in very helpful ways.*
>
> *Thanks, chaps. Not one of you, I think, will really agree with what I'm saying, and that's exactly as it should be.*

Inexhaustibility

30 March 2015

This is the first of four posts about key ideas from my book *The Shadow of Consciousness*. We start with the so-called Easy Problem, about how the human mind does its problem-solving, organism-guiding thing. If robots have Artificial Intelligence, we might call this the problem of Natural Intelligence.

I suggest that the real difficulty here is with what I call inexhaustible problems - a family of issues which includes non-computability but goes much wider. For the moment all I aim to do is establish that this is a meaningful group of problems and just suggest what the answer might be.

It's one of the ironies of the artificial intelligence project that Alan Turing both raised the flag for the charge and also set up one of the most serious obstacles. He declared that by the end of the twentieth century we should be able to speak of machines thinking without expecting to be contradicted; but he had already established, in his solution to the Halting Problem, that certain questions are unanswerable by the Universal Turing Machine and hence by the computers that approximate it. The human mind, though, is able to deal with these problems: so he seemed to have identified a wide gulf separating the human and computational performances he thought would come to be indistinguishable.

Turing himself said it was, in effect, merely an article of faith that the human mind did not ultimately, in respect of some problems, suffer the same kind of limitations as a computer; no-one had offered to prove it.

Non-computability, at any rate, was found to arise for a large set of problems; another classic example being the Tiling Problem. This relates to sets of tiles whose edges match, or fail to match, rather like dominoes. We can imagine that the tiles are square, with each edge a different colour, and that the rule is that wherever two edges meet, they must be the same colour. Certain sets of tiles will fit together in such a way that they will tile the plane: cover an infinite flat surface: others won't - after a while it becomes impossible to place another tile that matches. The problem is to determine whether any given set will tile the plane or not. This turns out unexpectedly to be a problem computers cannot answer. For certain sets of tiles, an algorithmic approach works fine; those that fail to tile the plain quite rapidly, and those that do so by forming repeating patterns like wallpaper. The fly in the ointment is that some elegant sets of tiles will cover the plane indefinitely, but only in a non-repeating, aperiodic way; when confronted with these, computational processes run on forever, unable to establish that the pattern will never begin repeating. Human beings, by resorting to other kinds of reasoning, can determine that these sets do indeed tile the plane.

Roger Penrose, who designed some examples of these aperiodic sets of tiles, also took up the implicit challenge thrown down by Turing, by attempting to prove that human thought is not affected by the limitations of computation. Penrose offered a proof that human mathematicians are not using a knowably sound algorithm to reach their conclusions. He did this by providing a

cunningly self-referential proposition stated in an arbitrary formal algebraic system; it can be shown that the proposition cannot be proved within the system, but it is also the case that human beings can see that in fact it must be true. Since all computers are running formal systems, they must be affected by this limitation, whereas human beings could perform the same extra-systemic reasoning whatever formal system was being used - so they cannot be affected in the same way.

Besides the fact that the human mind is not restricted to a formal system, Penrose established that it out-performs the machine by looking at meanings; the proposition in his proof is seen to be true because of what it says, not because of its formal syntactical properties.

Why is it that machines fail on these challenges, and how? In all these cases of non-computability the problem is that the machines start on processes which continue forever. The Turing Machine never halts, the tiling patterns never stop getting bigger - and indeed, in Penrose's proof the list of potential proofs which has to examined is similarly infinite. I think this rigorous kind of non-computability provides the sharpest, hardest-edged examples of a wider and vaguer family of problems arising from inexhaustibility.

A notable example of inexhaustibility in the wider sense is the Frame Problem, or at least its broader, philosophical version. In Dennett's classic exposition, a robot fails to notice an important fact; the trolley that carries its spare battery also bears a bomb. Pulling out the trolley has fatal consequences. The second version of the robot looks for things that might interfere with its safely regaining the battery, but is paralysed by the attempt to consider every logically possible deduction about the consequences of moving the trolley. A third robot is designed to identify only relevant events, but is equally paralysed by the task of considering the relevance of every possible deduction.

This problem is not so sharply defined as the Halting Problem or the Tiling Problem, but I think it's clear that there is some resemblance; here again computation fails when faced with an inexhaustible range of items. Combinatorial explosion is often invoked in these cases - the idea that when you begin looking at permutations of elements the number of combinations rises exponentially, too rapidly to cope with: that's not wrong, but I think the difficulty is deeper and arises earlier. Never mind combinations: even the initial range of possible elements for the AI to consider is already indefinably large.

Inexhaustible problems are not confined to AI. I think another example is Quine's indeterminacy of translation. Quine considered the challenge of interpreting an unknown language by relating the words used to the circumstances in which they were uttered. Roughly speaking, if the word "rabbit" is used exactly when a rabbit is visible, that's what it must mean; and through a series of observations we can learn the whole language. Unfortunately, it turns out that there is always an endless series of other things which the speaker might mean. Common sense easily rejects most of them - who on earth would talk about "sets of undetached rabbit parts"? - but what is the rigorous method that explains and justifies the conclusions that common sense reaches so easily? I said this was not an AI problem, but in a way it feels like one; arguably Quine was looking for the kind of method that could be turned into an algorithm.

In this case, we have another clue to what is going on with inexhaustible problems, albeit one which itself leads to a further problem. Quine assumed that the understanding of language was essentially a matter of decoding; we take the symbols and decode the meaning, process the meaning and recode the result into a new set of symbols. We know now that it doesn't really

work like that: human language rests very heavily on something quite different; the pragmatic reading of implicatures. We are able to understand other people because we assume they are telling us what is most relevant, and that grounds all kinds of conclusions which cannot be decoded from their words alone.

A final example of inexhaustibility requires us to tread in the footsteps of giants; David Hume, the Man Who Woke Kant, discovered a fundamental problem with cause and effect. How can we tell that A causes B? B consistently follows A, but so what? Things often follow other things for a while and then stop. The law of induction allows us to conclude that if B is regularly followed by A, we can conclude that it will go on doing so. But what justifies the law of induction? After all, many potential inductions are obviously false. Until quite recently a reasonable induction told us that Presidents of the United States were always white men.

Dennett pointed out that, although they are not the same, the Frame Problem and Hume's problem have a similar feel. They appear quite insoluble, yet ordinary human thought deals with them so easily it's sometimes hard to believe the problems are real. It's hard to escape the conclusion that the human mind has a faculty which deals with inexhaustible problems by some non-computational means. Over and over again we find that the human approach to these problems depends on a grasp of relevance or meaning; no algorithmic approach to either has been found.

So I think we need to recognise that this wider class of inexhaustible problem exists and has some common features. Common features suggest there might be a common solution, but what is it? Cutting to the chase, I think that in essence the special human faculty which lets us handle these problems so readily is simply recognition. Recognition is the ability to respond to entities in the world, and the ability to recognise larger entities as well as smaller ones within them opens the way to 'seeing where things are going' in a way that lets us deal with inexhaustible problems.

As I suggested recently, recognition is necessarily non-algorithmic. To apply rules, we need to have in mind the entities to which the rules apply. Unless these are just given, they have to be recognised. If recognition itself worked on the basis of rules, it would require us to identify a lower set of entities first - which again, could only be done by recognition, and so on indefinitely.

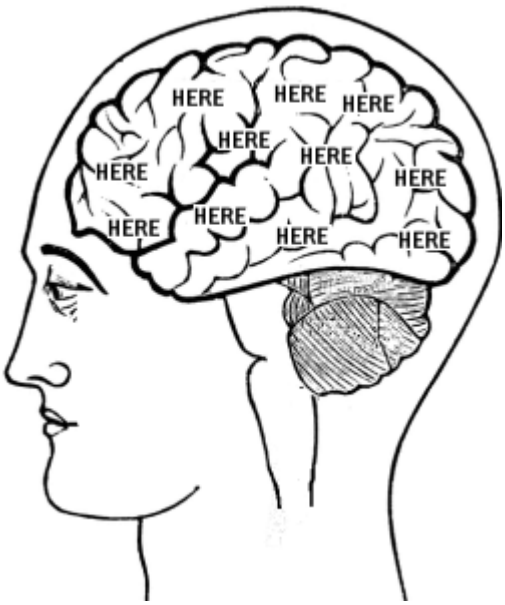
In our intellectual tradition, an informal basis like this feels unsatisfying, because we want proofs; we want something like Euclid, or like an Aristotelian syllogism. Hume took it that cause and effect could only be justified by either induction or deduction; what he really needed was recognition: recognition of the underlying entity of which both cause and effect are part. When we see that B is the result of A, we are really recognising that B is A a little later and transformed according to the laws of nature. Indeed, I'd argue that sometimes there is no transformation: the table sitting quietly over there is the cause of its own existence a few moments later.

As a matter of fact I claim that while induction relies directly on recognising underlying entities, even logical deduction is actually dependent on seeing the essential identity, under the laws of logic, of two propositions.

Maybe you're provisionally willing to entertain the idea that recognition might work as a basis for induction, sort of. But how, you ask, does recognition deal with all the other problems? I said that inexhaustible problems call for mastery of meaning and relevance: how does recognition account for those? I'll try to answer that in part 2.

That's You All Over

April 4 2015



An interesting study at Vanderbilt University (something not quite right about the brain picture on that page) suggests that consciousness is not narrowly localised within small regions of the cortex, but occurs when lots of connections to all regions are active. This is potentially of considerable significance, but some caution is appropriate.

The experiment asked subjects to report whether they could see a disc that flashed up only briefly, and how certain they were about it. Then it compared scans from occasions when awareness of the disc was clearly present or absent. The initial results provided the same kind of pattern we've become used to, in which small regions became active when awareness was present. Hypothetically these might be regions particularly devoted to disc detection; other studies in the past have found patterns and regions that appeared to be specific for individual objects, or even the faces of particular people.

Then, however, the team went on to assess connectedness and found that awareness was associated with many connections to all parts of the cortex. This might be taken to mean that while particular small bits of brain may have to do with particular things in the world, awareness itself is something the whole cortex does. This would be a very interesting result, congenial to some, and it would certainly affect the way we think about consciousness and its relation to the brain.

However, we shouldn't get too carried away too quickly. To begin with, the study was about awareness of a flashing disc; a legitimate example of a conscious state, but not a particularly complex one and not necessarily typical of distinctively human types of higher-level conscious activity. Second, I'm not remotely competent to make any technical judgement about the methods used to assess what connections were in place, but I'd guess there's a chance other teams in the field might have some criticisms.

Third, there seems to be scope for other interpretations of the results. At best we know that moments of disc awareness were correlated with moments of high connectedness. That might mean the connectedness caused or constituted the awareness, but it might also mean that it was just something that happens at the same time. Perhaps those narrow regions are still doing the real work: after all, when there's a key political debate the rest of the country connects up with it; but the debate still happens in a single chamber and would happen just the same if the wider connectivity failed. It might be that awareness gives a wide selection of other regions a chance to chip in, or to be activated in turn, but that that is not an essential feature of the experience of the disc.

For some people, the idea of consciousness bring radically decentralised will be unpalatable. To them, it's a functional matter which more or less has to happen in a defined area. OK, that area could be stretched out, but the idea that merely linking up disparate parts of the cortex could in

itself bring about a conscious state will seem too unlikely to be taken seriously. For others, who think the brain itself is too narrow an area to fully contain consciousness, the results will hardly seem to go far enough.

For myself, I feel some sympathy with the view expressed by Margaret Boden in this interview, where she speaks disparagingly of current neuroscience being mere ‘natural history’ - we just don’t have enough of a theory yet to know what we’re looking at. We’re still in the stage where we’re merely collecting facts, findings that will one day fit neatly into a proper theoretical framework, but at the moment don’t really prove or disprove any general hypotheses. To put it another way, we’re still collecting pieces of the jigsaw puzzle but we don’t have any idea what the picture is. When we spot that, then perhaps the pieces will all… connect.

Pointing

12 April 2015

This is the second of four posts about key ideas from my book The Shadow of Consciousness. This one looks at how the brain points at things, and how that could provide a basis for handling intentionality, meaning and relevance.

Intentionality is the quality of being about things, possessed by our thoughts, desires, beliefs and (clue's in the name) our intentions. In a slightly different way intentionality is also a property of books, symbols, signs and pointers. There are many theories out there about how it works; most, in my view, have some appeal, but none looks like the full story.

Several of the existing theories touch on a handy notion of natural meaning proposed by H.P.Grice. Natural meaning is essentially just the noticeable implication of things. Those spots mean measles; those massed dark clouds mean rain. If we regard this kind of 'meaning' as the wild, undeveloped form of intentionality we might be able to go on to suggest how the full-blown kind might be built out of it; how we get to non-natural meaning, the kind we generally use to communicate with and the kind most important to consciousness.

My proposal is that we regard natural meaning as a kind of pointing, and that pointing, in turn, is the recognition of a higher-level entity that links the pointer with the target. Seeing dark clouds and feeling raindrops on your head are two parts of the recognisable over-arching entity of a rain storm. Spots are just part of the larger entity of measles. So our basic ability to deal with meanings is simply a consequence of our ability to recognise things at different levels.

Looking at it that way, it's easy enough to see how we could build derived intentionality, the sort that words and symbols have; the difference is just that the higher-level entities we need to link everything up are artificial, supplied by convention or shared understanding: the words of a language, the conventions of a map. Clouds and water on my head are linked by the natural phenomenon of rain: the word 'rain' and water on my head are linked by the prodigious vocabulary table of the language. We can imagine how such conventions might grow up through something akin to a game of charades; I use a truncated version of a digging gesture to invite my neighbour to help with a hole: he gets it because he recognises that my hand movements could be part of the larger entity of digging. After a while the grunt I usually do at the same time becomes enough to convey the notion of digging.

External communication is useful, but this faculty of recognising wholes for parts and parts for wholes enables me to support more ambitious cognitive processes too, and make a bid for the original (aka 'intrinsic') intentionality that characterises my own thoughts, desires and beliefs. I start off with simple behaviour patterns in which recognising an object stimulates the appropriate behaviour; now I can put together much more complex stuff. I recognise an apple; but instead of just eating it, I recognise the higher entity of an apple tree; from there I recognise the long cycle of tree growth, then the early part in which a seed hits the ground; and from there

I recognise that the apple in my hand could yield the pips required, which are recognisably part of a planting operation I could undertake myself…

So I am able to respond, not just to immediate stimuli, but to think about future apples that don't even exist yet and shape my behaviour towards them. Plans that come out of this kind of process can properly be called intentional (I thought about what I was doing) and the fact that they seem to start with my thoughts, not simply with external stimuli, is what justifies our sense of responsibility and free will. In my example there's still an external apple that starts the chain of thought, but I could have been ruminating for hours and the actions that result might have no simple relationship to any recent external stimulus.

We can move thinks up another notch if I begin, as it were, to grunt internally. From the digging grunt and similar easy starts, I can put together a reasonable kind of language which not only works on my friends, but on me if I silently recognise the digging grunt and use it to pose to myself the concept of excavation.

There's more. In effect, when I think, I am moving through the forest of hierarchical relationships subserved by recognition. This forest has an interesting property. Although it is disorderly and extremely complex, it automatically arranges things so that things I perceive as connected in any way are indeed linked. This means it serves me as a kind of relevance space, where the things I may need to think about are naturally grouped and linked. This helps explain how the human brain is so good at dealing with the inexhaustible: it naturally (not infallibly) tends to keep the most salient things close.

In the end then, human style thought and human style consciousness (or at any rate the Easy Problem kind) seem to be a large and remarkably effective re-purposing of our basic faculty of recognition. By moving from parts to whole to other parts and then to other wholes, I can move through a conceptual space in a uniquely detached but effective way.

That's a very compressed version of thoughts that probably need a more gentle introduction, but I hope it makes some sense. On to haecceity!

Hard Problem Not Easy

19 April 2015

Antti Revonsuo has a two-headed paper in the latest JCS; at least it seems two-headed to me - he argues for two conclusions that seem to be only loosely related; both are to do with the Hard Problem, the question of how to explain the subjective aspect of experience.

The first is a view about possible solutions to the Hard Problem, and how it is situated strategically. Revonsuo concludes, basically, that the problem really is hard, which obviously comes as no great surprise in itself. His case is that the question of consciousness is properly a question for cognitive neuroscience, and that equally cognitive neuroscience has already committed itself to owning the problem: but at present no path from neural mechanisms up to conscious experience seems at all viable. A good deal of work has been done on the neural correlates of consciousness, but even if they could be fully straightened out it remains largely unclear how they are to furnish any kind of explanation of subjective experience.

The gist of that is probably right, but some of the details seem open to challenge. It's not at all clear to me that consciousness is owned by cognitive neuroscience; rather, the usual view is that it's an intensely inter-disciplinary problem; indeed, that may well be part of the reason it's so duffucult to get anywhere. Second, it's not at all that clear how strongly committed cognitive neuroscience is to the Hard Problem. Consciousness, fair enough; consciousness is indeed irretrievably one of the areas addressed by cognitive neuroscience. But consciousness is a many-splendoured thing, and I think cognitive neuroscientists still have the option of ignoring or being sceptical about some of the fancier varieties, especially certain conceptions of the phenomenal experience which is the subject of the Hard Problem. It seems reasonable enough that you might study consciousness in the Easy Problem sense - the state of being conscious rather than unconscious, we might say - without being committed to a belief in ineffable qualia - let alone to providing a neurological explanation of them.

The second conclusion is about extended consciousness; theories that suggest conscious states are not simply states of the brain but are partly made up of elements beyond our skull and our skin. These theories too, it seems, are not going to give us a quick answer in Revonsuo's opinion - or perhaps any answer. Revonsuo invokes the counter example of dreams. During dreams, we appear to be having conscious experiences; yet the difference between a dream state and an unconscious state may be confined to the brain; in every other respect our physical situation may be identical. This looks like strong evidence that consciousness is attributable to brain states alone.

Once, Revonsuo acknowledges, it was possible to doubt whether dreams were really experiences; it could have been that they were false memories generated only at the moment of awakening; but he holds that research over recent years has eliminated this possibility, establishing that dreams happen over time, more or less as they seem to.

The use of dreams in this context is not a new tactic, and Revonsuo quotes Alva Noë's counter-argument, which consists of three claims intended to undermine the relevance of dreams; first,

dream experiences are less rich and stable than normal conscious experiences; second, dream seeing is not real seeing; and third, all dream experiences depend on prior real experiences. Revonsuo more or less gives a flat denial of the first, suggesting that the evidence is thin to non-existent: Noë just hasn't cited enough evidence. He thinks the second counter-argument just presupposes that experiences without external content are not real experiences, which is question-begging. Just because I'm seeing a dreamed object, does that mean I'm not really seeing? On the third point he has two counter arguments. Even if all dreams recall earlier waking experiences, they are still live experiences in themselves; they're not just empty recall - but in any case, that isn't true; people who are congenitally paraplegic have dreams of walking, for example.

I think Revonsuo is basically right, but I'm not sure he has absolutely vanquished the extended mind. For his dream argument to be a real clincher, the brain state of dreaming of seeing a sheep and the brain state of actually seeing a sheep have to be completely identical, or rather, potentially identical. This is quite a strong claim to make, and whatever the state of the academic evidence, I'm not sure how well it stands up to introspective examination. We know that we often take dreams to be real when we are having them, and in fact do not always or even generally realise that a dream is a dream: but looking back on it, isn't there a difference of quality between dream states and waking states? I'm strongly tempted to think that while seeing a sheep is just seeing a sheep, the corresponding dream is about seeing a sheep, a little like seeing a film, one level higher in abstraction. But perhaps that's just my dreams?

Haecceity

26 April 2015

This is the third in a series of four posts about key ideas from my book The Shadow of Consciousness; this one is about haecceity, or to coin a plainer term, thisness. There are strong links with the subject of the final post, which will be that ultimate mystery, reality.

Haecceity is my explanation for the oddity of subjective experience. A whole set of strange stories are supposed to persuade us that there is something in subjective experience which is inexpressible, outside of physics, and yet utterly vivid and undeniable. It's about my inward experience of blue, which I can never prove is the same as yours; about what it is like to see red.

One of the best known thought experiments on this topic is the story of Mary the Colour Scientist. She has never seen colour, but knows everything there is to know about colour vision; when she sees a red rose for the first time, does she come to know something new? The presumed answer is yes: she now knows what it is like to see red things.

Another celebrated case asks whether I could have a 'zombie' twin, identical to me in every physical respect, who did not have these purely subjective aspects of experience – which are known as 'qualia', by the way. We're allowed to be unsure whether zombie twin is possible, but expected to agree that he is at least conceivable; and that that's enough to establish that there really is something extra going on, over and above the physics.

Most people, I think, accept that qualia do exist and do raise a problem, though some sceptics denounce the entire topic as more or less irretrievable nonsense. Qualia are certainly very odd; they have no causal effects, so nothing we say about them was caused by them: and they cannot be directly described. What we invariably have to do is refer to them by an objective counterpart: so we speak of the quale of hearing middle C, though middle C is in itself an irreproachably physical, describable thing (identifying the precisely correct physical counterpart for colour vision is actually rather complex, though I don't think anyone denies that you can give a full physical account of colour vision).

I suggest we can draw two tentative conclusions about qualia. First, knowledge of qualia is like knowledge of riding a bike: it cannot be transferred in words. I can talk until I'm blue in the face about bike riding, and it may help a little, but in the end to get that knowledge you have to get on a bike. That's because for bike riding it's your muscles and some non-talking parts of your brain that need to learn about it; it's a skill. We can't say the same about qualia because experiencing them is not a skill we need to learn; but there is perhaps a common factor; you have to have really done it, you have to have been there.

Second, we cannot say anything about qualia except through their objective counterparts. This leaves a mystery about how many qualia there are. Is there a quale of scarlet and a quale of crimson? An indefinite number of red qualia? We can't say, and since all hypotheses about the

number of qualia are equally good, we ought to choose the least expensive under the terms of Occam's Razor; the one with the fewest entities. It would follow from that that there is really only one universal quale; it provides the vivid liveliness while the objective aspects of the experience provide all the content.

So we have two provisional conclusions: all qualia are really the same thing conditioned differently by the objective features of the experience; and to know qualia you have to have 'been there', to have had real experience. I think it follows naturally from these two premises that qualia simply represent the particularity of experience; its haecceity. The aspect of experience which is not accounted for by any theory, including the theories of physics, is simply the actuality of experience. This is no discredit to theory: it is by definition about the general and the abstract and cannot possibly include the particular reality of any specific experience.

Does this help us with those two famous thought experiments? In Mary's case it suggests that what she knows after seeing the rose is simply what a particular experience is like. That could never have been conveyed by theoretical knowledge. In the case of my zombie twin, the real turning point is when we're asked to think whether he is conceivable; that transfers discussion to a conceptual, theoretical plane on which it is natural to suppose nothing has particularity.

Finally, I think this view explains why qualia are ineffable, why we can't say anything directly about them. All speech is, as it were, second order: it's about experiences, not the described experience itself. When we think of any objective aspect, we summon up the appropriate concepts and put them over in words; but when we attempt to convey the haecceity of an experience it drops out as soon as we move to a conceptual level. Description, for once, cannot capture what we want to convey.

There's nothing in all this that suggests anything wrong or incomplete about physics; no need for any dualism or magic realm. In a lot of ways this is simply the sceptical case approached more cautiously and from a different angle. It does leave us with some mystery though: what is it for something to be particular; what is the nature of particularity? We've already said we can't describe it effectively or reduce it theoretically, but surely there must be something we can do to apprehend it better? This is the problem of reality...

[Many thanks to Sergio for the kind review. Many thanks also to the generous people who have given me good reviews on amazon.com; much appreciated!]

Mind Uploading: Does Speed Matter?

5 May 2015

Not according to Keith B. Wiley and Randal A.Koene. They contrast two different approaches to mind uploading: in the slow version neurons or some other tiny component are replaced one by one with a computational equivalent; in the quick, the brain is frozen, scanned, and reproduced in a suitable computational substrate. Many people feel that the slow way is more likely to preserve personal identity across the upload, but Wiley and Koene argue that it makes no difference. Why does the neuron replacement have to be slow? Do we have to wait a day between each neuron switch? hard to see why - why not just do the switch as quickly as feasible? Putting aside practical issues (we have to do that a lot in this discussion), why not throw a single switch that replaces all the neurons in one go? Then if we accept that, how is it different from a destructive scan followed immediately by creation of the computational equivalent (which, if we like, can be exactly the same as the replacement we would have arrived at by the other method)? If we insist on a difference, argue Wiley and Koene, then somewhere along the spectrum of increasing speed there must be a place where preservation of identity switches abruptly to loss of identity; this is quite implausible and there are no reasonable arguments that suggest where this maximum speed should occur.

One argument for the difference comes from non-destructive scanning. Wiley and Koene assume that the scanning process in the rapid transfer is destructive; but if it were not, the original brain would continue on its way, and there would be two versions of the original person. Equally, once the scanning is complete there seems to be no reason why multiple new copies could not be generated. How can identity be preserved if we end up with multiple versions of the original? Wiley and Koene believe that once we venture into this area we need to expand our concept of identity to include the possibility of a single identity splitting, so for them this is not a fatal objection.

Perhaps the problem is not so much the speed in itself as the physical separation? In the fast version the copy is created some distance away from the original whereas in gradual replacement the new version occupies essentially the same space as the original - might it be this physical difference which gives rise to differing intuitions about the two methods? Wiley and Koene argue that even in the case of gradual replacement, there is a physical shift. The replacement neuron cannot occupy exactly the same space as the one it is to replace, at least not at the moment of transfer. The spatial difference may be a matter of microns rather then metres, but here again, why should that make a difference? As with speed, are going to fix on some arbitrary limit where the identity ceases to be preserved, and why should that happen?

I think Wiley and Koene don't do full justice to the objection here. I don't think it really rests on physical separation; it implicitly rests on continuity. Wiley and Koene dismiss the idea that a continuous stream of consciousness is required to preserve identity, but it can be defended. It rests on the idea that personal identity resides not in the data or the function in general, but a specific instance in particular. We might say that I as a person am not equivalent to SimPerson V

- I am equivalent to a particular game of SimPerson V, played on a particular occasion. If I reproduce that game exactly on another occasion, it isn't me, it's a twin.

Now the gradual replacement scenario arguably maintains that kind of identity. The new, artificial neurons enter an ongoing live process and become part of it, whereas in the scan and create process the brain is necessarily stopped, translated into data, and then recreated. It's neither the speed nor the physical separation that disrupts the preservation of the identity: it's the interruption.

Can that be right though - is merely stopping someone enough to disrupt their identity? What if I were literally frozen in some icy accident, so that my brain flat-lined; and then restored and brought back to life. Are we forced to say that the person after the freezing is a twin, not the same? That doesn't seem right. Perhaps brute physical continuity has some role to play after all; perhaps the fact that when I'm frozen and brought back it's the same neurons that are firing helps somehow to sustain the identity of the process over the stoppage and preserve my identity?

Or perhaps Wiley and Koene are right after all?

Reality

11 May 2015

This is the last of four posts about key ideas from my book The Shadow of Consciousness. This time the subject is reality.

Last time I suggested that qualia - the subjective aspect of experiences that gives them their what-it-is-like quality - are just the particularity, or haecceity, of real experiences. It is like something to see that red because you're really seeing it; you're not just understanding the theory. So we could say qualia are just the reality of experience. No mystery about it after all.

Except of course there is a mystery - what is reality? There's something oddly arbitrary about reality; some things are real, others are not. That cake on the table in front of me; it could be real as far as you know; or it could indeed be that the cake is a lie. The number 47, though, is quite different; you don't need to check the table or any location; you don't need to look for an example or count to fifty; it couldn't have been the case that there was no number 47. Things that are real in the sense we need for haecceity seem to depend on events for their reality. I will borrow some terminology from Meinong and call that dependent or contingent kind of reality existence, while what the number 47 has got is subsistence.

What is existence, then? Things that exist depend on events, I suggested; if I made a cake and put it in the table, it exists; if no-one did that, it doesn't. Real things are part of a matrix of cause and effect, a matrix we could call history. Everything real has to have causes and effects. We can prove that perhaps, by considering the cake's continuing existence. It exists now because it existed a moment ago; if it had no causal effects, it wouldn't be able to cause its own future reality, and it wouldn't be here. If it wasn't here, then it couldn't have had preceding causes, so it didn't exist in the past either. Ergo, things without causal effects don't exist.

Now that's interesting because of course, one of the difficult things about qualia is that they apparently can't have causal effects. If so, I seem to have accidentally proved that they don't exist! I think things get unavoidably complex here. What I think is going on is that qualia in general, the having of a subjective side, is bestowed on things by being real, and that reality means causal efficacy. However, particular qualia are determined by the objective physical aspects of things; and it's those that give specific causal powers. It looks to us as if qualia have no causal effects because all the particular causal powers have been accounted for in the objective physical account. There seems to be no role for qualia. What we miss is that without reality nothing has causal powers at all.

Let's digress slightly to consider yet again my zombie twin. He's exactly like me, except that he has no qualia, and that is supposed to show that qualia are over and above the account given by physics. Now according to me that is actually not possible, because if my zombie twin is real, and physically just the same, he must end up with the same qualia. However, if we doubt this possibility, David Chalmers and others invite us at least to accept that he is conceivable. Now

we might feel that whether we can or can't conceive of a thing is a poor indicator of anything, but leaving that aside I think the invitation to consider the zombie twin's conceivability draws us towards thinking of a conceptual twin rather than a real one. Conceptual twins - imaginary, counterfactual, or nonexistent ones - merely subsist; they are not real and so the issue of qualia does not arise. The fact that imaginary twins lack qqualia doesn't prove what it was meant to; properly understood it just shows that qualia are an aspect of real experience.

Anyway, are we comfortable with the idea of reality? Not really, because the buzzing complexity and arbitrariness of real things seems to demand an explanation. If I'm right about all real things necessarily being part of a causal matrix, they are in the end all part of one vast entity whose curious firm should somehow be explicable.

Alas, it isn't. We have two ways of explaining things. One is pure reason: we might be able to deduce the real world from first principles and show that it is logically necessary. Unfortunately pure reason alone is very bad at giving us details of reality; it deals only with Platonic, theoretical entities which subsist but do not exist. To tell us anything about reality it must at least be given a few real facts to work on; but when we're trying to account for reality as a whole that's just what we can't provide.

The other kind of explanation we can give is empirical; we can research reality itself scientifically and draw conclusions. But empirical explanations operate only within the causal matrix; they explain one state of affairs in terms of another, usually earlier one. It's not possible to account for reality itself this way.

It looks then, as if reality is doomed to remain at least somewhat mysterious, unless we somehow find a third way, neither empirical nor rational.

A rather downbeat note to end on, but sincere thanks to all those who have helped make the discussion so interesting so far.

Alters Of The Universe

17 May 2015

Bernardo Kastrup has some marvellous invective against AI engineers in this piece...¶

> *The computer engineer's dream of birthing a conscious child into the world without the messiness and fragility of life is an infantile delusion; a confused, partial, distorted projection of archetypal images and drives. It is the expression of the male's hidden aspiration for the female's divine power of creation. It represents a confused attempt to transcend the deep-seated fear of one's own nature as a living, breathing entity condemned to death from birth. It embodies a misguided and utterly useless search for the eternal, motivated only by one's amnesia of one's own true nature. The fable of artificial consciousness is the imaginary band-aid sought to cover the engineer's wound of ignorance.*
>
> *I have been this engineer.*

I think it's untrue, but you don't have to share the sentiment to appreciate the splendid rhetoric.

Kastrup distinguishes intelligence, which is a legitimate matter of inputs, outputs and the functions that connect them, from consciousness, the true what-it-is likeness of subjectivity. In essence he just doesn't see how setting up functions in a machine can ever touch the latter.

Not that Kastrup has a closed mind, he speaks approvingly of Pentti Haikonen's proposed architecture; he just doesn't think it works. As Kastrup sees it Haikonen's network merely gathers together sparks of consciousness: it then does a plausible job of bringing them together to form more complex kinds of cognition, but in Kastrup's eye it assumes that consciousness is there to be gathered in the first place: that it exists out there in tiny parcels amendable to this kind of treatment; there is, he thinks, absolutely no reason to think that this kind of panpsychism is true: no reason to think that rocks or drops of water have any kind of conscious experience at all.

I don't know whether that is the right way to construe Haikonen's project (I doubt whether gathering experiential sparks is exactly what Haikonen supposed he was about). Interestingly, though Kastrup is against the normal kind of panpsychism (if the concept of 'normal panpsychism' is admissible), his own view is essentially a more unusual kind.

Kastrup considers that we're dealing with two aspects here; internal and external; our minds have both; the external is objective, the internal represents subjectivity. Why wouldn't the world also have these two aspects? (Actually it's hard to say why anything should have them, and we may suspect that by taking it as a given we're in danger of smuggling half the mystery out of the problem, but let's let that pass.) Kastrup takes it as natural to conclude that the world as a whole must indeed have the two aspects (I think at this point he may have inadvertently proved the existence of God, which is regrettable, but again let's go with it for now); but not parts of the world. The brain, we know, has experience, but the groups of neurons that make it up do not (do

we actually know that?); it follows that while the world as a whole has an internal aspect, objects or entities within it generally do not.

Yet of course, the brain manages it, which must surely be something to do with the structure of the brain? May we not suspect that whatever it is that allows the brain to have an internal aspect, a machine could in principle have it too? I don't think Kastrup engages sufficiently with this objection; his view seems to be that metabolism is essential, though why that should be, and why machines can't have some form of metabolism, we don't know.

The argument, then, doesn't seem convincing, but it must be granted that Kastrup has an original and striking vision: our consciousnesses, he suggests, are essentially like the 'alters' of Dissociative Identity Disorder, better known as Multiple Personality, in which several different people seem to inhabit a single human being. We are, he says, like the accidental alternate identities of the Universe (I think you could say, of God, though Kastrup clearly doesn't want to).

Again, I don't think at all that that is the case, but it is a great idea. It is probable that in his book-length treatments of these ideas Kastrup makes a stronger case than I have given him credit for above, but I do admire the originality of his thinking, and the clarity and force with which he expresses it.

Ex Machina

25 May 2015

I finally saw Ex Machina, which everyone has been telling me is the first film about artificial intelligence you can take seriously. Competition in that area is not intense, of course: many films about robots and conscious computers are either deliberately absurd or treat the robot as simply another kind of monster. Even the ones that cast the robots as characters in a serious drama are essentially uninterested in their special nature and use them as another kind of human, or at best to make points about humanity. But yes: this one has a pretty good grasp of the issues about machine consciousness and even presents some of them quite well, up to and including Mary the Colour Scientist. (Spoilers follow.)

If you haven't seen it (and I do recommend it), the core of the story is a series of conversations between Caleb, a bright but naive young coder, and Ava, a very female robot. Caleb has been told by Nathan, Ava's billionaire genius creator, that these conversations are a sort of variant Turing Test. Of course in the original test the AI was a distant box of electronics: here she's a very present and superficially accurate facsimile of a woman. (What Nathan has achieved with her brain is arguably overshadowed by the incredible engineering feat of the rest of her body. Her limbs achieve wonderful fluidity and power of movement, yet they are transparent and we can see that it's all achieved with something not much bigger than a large electric cable. Her innards are so economical there's room inside for elegant empty spaces and decorative lights. At one point Nathan is inevitably likened to God, but on anthropomorph engineering design he seems to leave the old man way behind.)

Why does she have gender? Caleb asks, and is told that without sex humans would never have evolved consciousness; it's a key motive, and hell, it's fun. In story terms making Ava female perhaps alludes to the origin of the Turing Test in the Imitation Game, which was a rather camp pastime about pretending to be female played by Turing and his friends. There are many echoes and archetypes in the film; Bluebeard, Pygmalion, Eros and Psyche to name but three; all of these require that Ava be female. If I were a Jungian I'd make something of that.

There's another overt plot reason, though; this isn't really a test to determine whether Ava is conscious, it's about whether she can seduce Caleb into helping her escape. Caleb is a naive girl-friendless orphan; she has been designed not just as a female but as a match for Caleb's preferred porn models (as revealed in the search engine data Nathan uses as his personal research facility - he designed the search engine after all). What a refined young Caleb must be if his choice of porn revolves around girls with attractive faces (on second thoughts, let's not go there).

We might suspect that this test is not really telling us about Ava, but about Caleb. That, however, is arguably true of the original Turing Test too. No output from the machine can prove consciousness; the most brilliant ones might be the result of clever tricks and good luck. Equally, no output can prove the absence of consciousness. I've thought of entering the

Loebner prize with Swearbot, which merely replies to all input with “Shut the fuck up” - this vividly resembles a human being of my acquaintance.

There is no doubt that the human brain is heavily biased in favour of recognising things as human. We see faces in random patterns and on machines; we talk to our cars and attribute attitudes to plants. No doubt this predisposition made sense when human beings were evolving. Back then, the chances of coming across anything that resembled a human being without it being one were low, and given that an unrecognised human might be a deadly foe or a rare mating opportunity the penalties for missing a real one far outweighed those for jumping at shadows or funny-shaped trees now and then.

Given all that, setting yourself the task of getting a lonely young human male romantically interested in something not strictly human is perhaps setting the bar a bit low. Naked shop-window dummies have pulled off this feat. If I did some reprogramming so that the standard utterance was a little dumb-blonde laugh followed by “Let’s have fun!” I think even Swearbot would be in with a chance.

I think the truth is that to have any confidence about an entity being conscious, we really need to know something about how it works. For human beings the necessary minimum is supplied by the fact that other people are constituted much the same way as I am and had similar origins, so even though I don’t know how I work, it’s reasonable to assume that they are similar. We can’t generally have that confidence with a machine, so we really need to know both roughly how it works and - bit of a stumper this - how consciousness works.

Ex Machina doesn’t have any real answers on this, and indeed doesn’t really seek to go much beyond the ground that’s already been explored. To expect more would probably be quite unreasonable; it means though, that things are necessarily left rather ambiguous.

It’s a shame in a way that Ava resembles a real woman so strongly. She wants to be free (why would an AI care, and why wouldn’t it fear the outside world as much as desire it?), she resents her powerlessness; she plans sensibly and even manipulatively and carries on quite normal conversations. I think there is some promising scope for a writer in the oddities that a genuinely conscious AI’s assumptions and reasoning would surely betray, but it’s rarely exploited; to be fair Ex Machina has the odd shot, notably Ava’s wish to visit a busy traffic intersection, which she conjectures would be particularly interesting; but mostly she talks like a clever woman in a cell. (Actually too clever: in that respect not too human).

At the end I was left still in doubt. Was the take-away that we’d better start thinking about treating AIs with the decent respect due to a conscious being? Or was it that we need to be wary of being taken in by robots that seem human, and even sexy, but in truth are are dark and dead inside?

Six Consciousnesses

1 June 2015

Over at Brains Blog Uriah Kriegel has been doing a series of posts (starting here) on some themes from his book The Varieties of Consciousness, and in particular his identification of six kinds of phenomenology.

I haven't read the book (yet) and there may be important bits missing from the necessarily brief account given in the blog posts, but it looks very interesting. Kriegel's starting point is that we probably launch into explaining consciousness too quickly, and would do well to spend a bit more time describing it first. There's a lot of truth in that; consciousness is an extraordinarily complex and elusive business, yet phenomenology remains in a pretty underdeveloped state. However, in philosophy the borderline between describing and explaining is fuzzy; if you're describing owls you can rely on your audience knowing about wings and beaks and colouration; in philosophy it may be impossible to describe what you're getting at without hacking out some basic concepts which can hardly help but be explanatory. With that caveat, it's a worthy project.

Part of the difficulty of exploring phenomenology may come from the difficulty of reconciling differences in the experiences of different reporters. Introspection, the process of examining our own experience, is irremediably private, and if your conclusions are different from mine, there's very little we can do about it other than shout at each other. Some have also taken the view that introspection is radically unreliable in any case, a task like trying to watch the back of your own head; the Behaviourists, of course, concluded that it was a waste of time talking about the contents of consciousness at all: a view which hasn't completely disappeared.

Kriegel defends introspection, albeit in a slightly half-hearted way. He rightly points out that we've tacitly relied on it to support all the discoveries and theorising which has been accomplished in recent decades. He accepts that we cannot any longer regard it as infallible, but he's content if it can be regarded as more likely right than wrong.

With this mild war-cry, we set off on the exploration. There are lots of ways we can analyse consciousness, but what Kriegel sets out to do is find the varieties of phenomenal experience. He's come up with six, but it's a tentative haul and he's not asserting that this is necessarily the full set. The first two phenomenologies, taken as already established, are the perceptual and the algedonic (pleasure/pain); to these Kriegel adds: cognitive phenomenology, "conative" phenomenology (to do with action and intention), the phenomenology of entertaining an idea or a proposition (perhaps we could call it 'considerative', though Kriegel doesn't), and the phenomenology of imagination.

The idea that there is conative phenomenology is a sort of cousin of the idea of an 'executive quale' which I have espoused: it means there is something it is like to desire, to decide, and to intend. Kriegel doesn't spend any real effort on defending the idea that these things have phenomenology at all, though it seems to me (introspectively!) that sometimes they do and

sometimes they don't. What he is mainly concerned to do is establish the distinction between belief and desire. In non-phenomenal terms these two are sort of staples of the study of intentionality: Bel and Des, the old couple. One way of understanding the difference is in terms of 'direction of fit', a concept that goes back to J.L. Austin. What this means is that if there's a discrepancy between your beliefs and the world, then you'd better change your beliefs. If there's a discrepancy between your desires and the world, you try to change the world (usually: I think Andy Warhol for one suggested that learning to like what was available was a better strategy, thereby unexpectedly falling into a kind of agreement with some religious traditions that value acceptance and submission to the Divine Will).

Kriegel, anyway, takes a different direction, characterising the difference in terms of phenomenal presentation. What we desire is presented to us as good; what we believe is presented as true. This approach opens the way to a distinction between a desire and a decision: a desire is conditional (if circumstances allow, you'll eat an ice-cream) whereas a decision is categorical (you're going to eat an ice-cream). This all works quite well and establishes an approach which can handily be applied to other examples; if we find that there's presentation-as-something different going on we should suspect a unique phenomenology. (Are we perhaps straying here into something explanatory instead of merely descriptive? I don't think it matters.) I wonder a bit about whether things we desire are presented to us as good. I think I desire some things that don't seem good at all except in the sense that they seem desirable. That's not much help, because if we're reduced to saying that when I desire something it is presented to me as desirable we're not saying all that much, especially since the idea of presentation is not particularly clarified. I have no doubt that issues like this are explored more fully in the book.

Kriegel moves on to consider the case of emotion: does it have a unique and irreducible phenomenology? If something we love is presented to us as good, then we're back with the merely conative; and Kriegel doesn't think presentation as beautiful is going to work either (partly because of negative cases, though I don't see that as an insoluble probem myself; if we can have algedonia, the combined quality of pain or pleasure we can surely have an aesthetic quality that combines beauty and ugliness). In the end he suspects that emotion is about presentation as important, but he recognises that this could be seen as putting the cart before the horse; perhaps emotion directs our attention to things and what gets our attention seems to be important. Kriegel finds it impossible to decide whether emotion has an independent phenomenology and gives the decision by default in favour of the more parsimonious option, that it is reducible to other phenomenologies.¶On that, it may be that taking all emotion together was just too big a bite. It seems quite likely to me that different emotions might have different phenomenologies, and perhaps tackling it that way would yield more positive results.

Anyway, a refreshing look at consciousness.

The Stance Scanned

8 June 2015

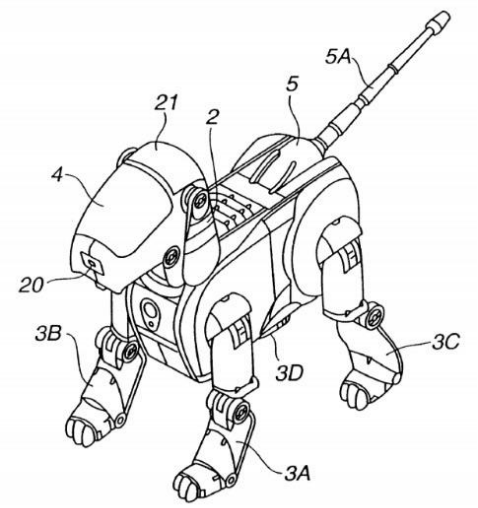
Dan Dennett famously based his view of consciousness on the intentional stance. According to him the attribution of intentions and other conscious states is a most effective explanatory strategy when applied to human beings, but that doesn't mean consciousness is a mysterious addition to physics. He compares the intentions we attribute to people with centres of gravity, which also help us work out how things will behave, but are clearly not a new set of real physical entities.¶Whether you like that idea or not, it's clear that the human brain is strongly predisposed towards attributing purposes and personality to things. Now a new study using FMRI provides evidence that even when the brain is ostensibly not doing anything, it is in effect ready to spot intentions.¶This is based on findings that similar regions of the brain are active both in a rest state and when making intentional (but not non-intentional) judgements, and that activity in the pre-frontal cortex of the kind observed when the brain is at rest is also associated with greater ease and efficiency in making intentional attributions.¶There's always some element of doubt about how ambitious we can be in interpreting what FMRI results are telling us, and so far as I can see it's possible in principle that if we had a more detailed picture than FMRI can provide we might see more significant differences between the rest state and the attribution of intentions; but the researchers cite evidence that supports the view that broad levels of activity are at least a significant indicator of general readiness.¶You could say that this tells us less about intentionality and more about the default state of the human mind. Even when at rest, on this showing, the brain is sort of looking out for purposeful events. In a way this supports the idea that the brain is never naturally quiet, and explains why truly emptying the mind for purposes of deep meditation and contemplation might require deliberate preparation and even certain mental disciplines.¶So far as consciousness itself is concerned, I think the findings lend more support to the idea that having 'theory of mind' is an essential part of having a mind: that is, that being able to understand the probable point of view and state of knowledge of other people is a key part of having full human-style consciousness yourself.¶There's obviously a bit of a danger of circularity there, and I've never been sure it's a danger that Dennett for one escapes. I don't know how you attribute intentions to people unless you already know what intentions are. The normal expectation would be that I can do that because I have direct knowledge of my own intentions, so all I need to do is hypothesise that someone is thinking the way I would think if I were in their shoes. In Dennett's theory, me having intentions is really just more attribution (albeit self-attribution), so we need some other account of how it all gets started (apparently the answer is that we assume optimal intentions in the light of assumed goals).¶Be that as it may, the idea that consciousness involves attributing conscious states to ourselves is one that has a wider appeal and it may shed a slightly different light on the new findings. It might be that the base activity identified by the study is not so much a readiness to attribute intentions, but a continuous second-order contemplation of our own intentions, and

an essential part of normal consciousness. This wouldn't mean the paper's conclusions are wrong, but it would suggest that it's consciousness itself that makes us more ready to attribute intentions.

Hard to test that one because unconscious patients would not make co-operative subjects.

The Dog It Was

21 June 2015



The recent short NYT series on robots has a dying fall. The articles were framed as an investigation of how robots are poised to change our world, but the last piece is about the obsolescence of the Aibo, Sony's robot dog. Once apparently poised to change our world, the Aibo is no longer made and now Sony will no longer supply spare parts, meaning the remaining machines will gradually cease to function.

There is perhaps a message here about the over-selling and under-performance of many ambitious AI projects, but the piece focuses instead on the emotional impact that the 'death' of the robot dogs will have on some fond users. The suggestion is that the relationship these owners have with their Aibo is as strong as the one you might have with a real dog. Real dogs die, of course, so though it may be sad, that's nothing new. Perhaps the fact that the Aibos are 'dying' as the result of a corporate decision, and could in principle have been immortal makes it worse? Actually I don't know why Sony or some third party entrepreneur doesn't offer a program to virtualise your Aibo, uploading it into software where you can join it after the Singularity (I don't think there would really be anything to upload, but hey…).

On the face of it, the idea of having a real emotional relationship with an Aibo is a little disturbing. Aibos are neat pieces of kit, designed to display 'emotional' behaviour, but they are not that complex (many orders of magnitude less complex than a dog, surely), and I don't think there is any suggestion that they have any real awareness or feelings (even if you think thermostats have vestigial consciousness, I don't think an Aibo would score much higher. If people can have fully developed feelings for these machines, it strongly suggests that their feelings for real dogs have nothing to do with the dog's actual mind. The relationship is essentially one-sided; the real dog provides engaging behaviour, but real empathy is entirely absent.

More alarming, it might be thought to imply that human relationships are basically the same. Our friends, our loved ones, provide stimuli which tickle us the right way; we enjoy a happy congruence of behaviour patterns, but there is no meeting of minds, no true understanding. What's love got to do with it, indeed?

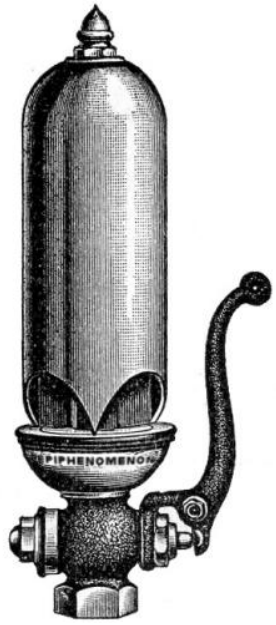
Perhaps we can hope that Aibo love is actually quite distinct from dog love. The people featured in the NYT video are Japanese, and it is often said that Japanese culture is less rigid about the distinction between animate and inanimate than western ideas. In Christianity, material things lack souls and any object that behaves as if it had one may be possessed or enchanted in ways that are likely to be unnatural and evil. In Shinto, the concept of kami extends to anything important or salient, so there is nothing unnatural or threatening about robots. But while that might validate the idea of an Aibo funeral, it does not precisely equate Aibos and real dogs.¶In fact, some of the people in the video seem mainly interested in posing their Aibos for amusing

pictures or video, something they could do just as well with deactivated puppets. Perhaps in reality Japanese culture is merely more relaxed about adults amusing themselves with toys?

Be that as it may, it seems that for now the era of robot dogs is already over.

Unimportant Consciousness

5 July 2015



An interesting paper in Behavioural and Brain Sciences from Morsella, Godwin, Jantz, Krieger, and Gazzaley.

It's quite a complex, thoughtful paper, but the gist is clearly that consciousness doesn't really do that much. The authors take the view that many functions generally regarded as conscious are in fact automatic and pre- or un-conscious: what consciousness hosts is not the process but the results. It looks to consciousness as though it's doing the work, but really it isn't.

In itself this is not a new view, of course. We've heard of other theories that base their interpretation on the idea that consciousness only deals with small nuggets of information fed to it by unconscious processes. Indeed, as the authors acknowledge, some take the view that consciousness does nothing at all: that it is an epiphenomenon, a causal dead end, adding no more to human behaviour than the whistle adds to the locomotive.

Morsella et al don't go that far. In their view we're lacking a clear idea of the prime function of consciousness; their Passive Frame Theory holds that the function is to constrain and direct skeletal muscle output, thereby yielding adaptive behaviour. I'd have thought quite a lot of unconscious processes, even simple reflexes, could be said to do that too; philosophically I think we'd look for a bit more clarity about the characteristic ways in which consciousness as opposed to instinct or other unconscious urges influence behaviour; but perhaps I'm nit-picking.

The authors certainly offer explanation as to what consciousness does. In their view, well-developed packages are delivered to consciousness from various unconscious functions. In consciousness these form a kind of combinatorial jigsaw, very regularly refreshed in a conscious/unconscious cycle; the key thing is that these packages are encapsulated and cannot influence each other. This is what distinguishes the theory from the widely popular idea of a Global Workspace, originated by Bernard Baars; no work is done on the conscious contents while they're there. they just sit until refreshed or replaced.

The idea of encapsulation is made plausible by various examples. When we recognise an illusion, we don't stop experiencing it; when we choose not to eat, we don't stop feeling hungry, and so on. It's clearly the case that sometimes this happens, but can we say that there are really no cases where one input alters our conscious perception of another? I suspect that any examples we might come up with would be deemed by Morsella et al to occur in pre-conscious processing and only seem to happen in consciousness: the danger with that is that the authors might end up simply disqualifying all counter-examples and thereby rendering their thesis unfalsifiable. It would help if we could have a sharper criterion of what is, and isn't, within consciousness.

As I say, the authors do hold that consciousness influences behaviour, though not by its own functioning; instead it does so by, in effect, feeding other unconscious functions. An analogy with the internet is offered: the net facilitates all kind of functions; auctions, research, social

interaction, filling the world with cat pictures, and so on; but it would be quite right to say that in itself it does any of these things.

That's OK, but it seems to delegate an awful lot of things that we might have regarded as conscious cognitive activity to these later unconscious functions, and it would help to have more of an account of how they do their thing and how consciousness contrives to facilitate. It seems it merely brings things together, but how does that help? If they're side by side but otherwise unprocessed, I'm not sure what the value of merely juxtaposing them (in some sense which is itself a little unclear) amounts to.

So I think there's more to do if Passive Frame Theory is going to be a success; but it's an interesting start.

Selfless Transfers

13 July 2015

The new film Self/less is based around the transfer of consciousness. An old man buys a new body to transfer into, and then finds that contrary to what he was told, it wasn't grown specially: there was an original tenant who moreover, isn't really gone. I understand that this is not a film that probes the metaphysics of the issue very deeply; it's more about fight scenes; but the interesting thing is how readily we accept the idea of transferred consciousness.

In fact, it's not at all a new idea; if memory serves, H.G.Wells wrote a short story on a similar theme; a fit young man with no family is approached by a rich old man in poor health who apparently wants to leave him all his fortune; then he finds himself transferred unwillingly to the collapsing ancient body and the old man making off in his fresh young one. In Wells' version the twist is that the old man gets killed in a chance traffic accident, thereby dying before his old body does anyway.

The thing is, how could a transfer possibly work? In Wells' story it's apparently done with drugs, which is mysterious; more normally there's some kind of brain-transfer helmet thing. It's pretty much as though all you needed to do was run an EEG and then reverse the polarity. That makes no sense. I mean, scanning the brain in sufficient detail is mind-boggling to begin with, but the idea that you could then use much the same equipment to change the content of the mind is in another league of bogglement. Weather satellites record the meteorology of the world, but you cannot use them to reset it. This is why uploading your mind to a computer, while highly problematic, is far easier to entertain than transferring it to another biological body.

The big problem is that part of the content of the brain is, in effect, structural. It depends on which neurons are attached to which (and for that matter, which neurons there are), and on the strength and nature of that linkage. It's true that neural activity is important too, and we can stimulate that electrically; even with induction gear that resembles the SF cliché: but we can't actually restructure the brain that way.

The intuition that transfers should be possible perhaps rests on an idea that the brain is, as it were, basically hardware, and consciousness is essentially software; but isn't really like that. You can't run one person's mind on another's brain.

There is in fact no reason to suppose that there's much of a read-across between brains: they may all be intractably unique. We know that there tends to be a similar regionalisation of functions in the brain, but there's no guarantee that your neural encoding of 'grandmother' resembles mine or is similarly placed. Worse, it's entirely possible that the 'meaning' of neural assemblages differs according to context and which other structures are connected, so that even if I could identify my 'grandmother' neurons, and graft them in in place of yours, they would have a different significance, or none.

Perhaps we need a more sophisticated and bespoke approach. First we thoroughly decipher both brains, and learn their own idiosyncratic encoding works. Then we work out a translator. This is a task of unimaginable complexity and particularity, but it's not obviously impossible in principle. I think it's likely that for each pair of brains you would need a unique translator: a universal one seems such an heroic aspiration that I really doubt its viability: a universal mind map would be an achievement of such interest and power that merely transferring minds would seem like time-wasting games by comparison.

I imagine that even once a translator had been achieved, it would normally only achieve partial success. There would be a limit to how far you could go with nano-bot microsurgery; and there might be certain inherent limitations. Certain ideas, certain memories, might just be impossible to accommodate in the new brain because of their incompatibility with structural or conceptual issues that were too deeply embedded to change; there would be certain limitations. The task you were undertaking would be like the job of turning Pride and Prejudice into Don Quixote simply by moving the words around and perhaps in a few key cases allowing yourself one or two anagrams: the result might be recognisable, but it wouldn't be perfect. The transfer recipient would believe themselves to be Transferee, but they would have strange memory gaps and certain cognitive deficits, perhaps not unlike Alzheimer's, as well as artefacts: little beliefs or tendencies that existed neither in Transferee or Recipient, but were generated incidentally through the process of reconstruction.

It's a much more shadowy and unappealing picture, and it makes rather clearer the real killer: that though Recipient might come to resemble Transferee, they wouldn't really be them.

In the end, we're not data, or a program; we're real and particular biological entities, and as such our ontology is radically different. I said above that the plausibility of transfers comes from thinking of consciousness as data, which I think is partly true: but perhaps there's something else going on here; a very old mental habit of thinking of the soul as detachable and transferable. This might be another case where optimists about the capacity of IT are unconsciously much less materialist than they think.

Slippery Humanity

26 July 2015

There were a number of reports recently that a robot had passed 'one of the tests for self-awareness'. They seem to stem mainly from this New Scientist piece (free registration may be required to see the whole thing, but honestly I'm not sure it's worth it). That in turn reported an experiment conducted by Selmer Bringsjord of Rensselaer, due to be presented at the Ro-Man conference in a month's time. The programme for the conference looks very interesting and the experiment is due to feature in a session on 'Real Robots That Pass Human Tests of Self Awareness'.

The claim is that Bringsjord's bot passed a form of the Wise Man test. The story behind the Wise Man test has three WMs tested by the king; he makes them wear hats which are either blue or white: they cannot see their own hat but can see both of the others. They're told that there is at least one blue hat, and that the test is fair; to be won by the first WM who correctly announces the colour of his own hat. There is a chain of logical reasoning which produces the right conclusion: we can cut to the chase by noticing that the test can't be fair unless all the hats are the same colour, because all other arrangements give one WM an advantage. Since at least one hat is blue, they all are.

You'll notice that this is essentially a test of logic, not self awareness. If solving the problem required being aware that you were one of the WMs then we who merely read about it wouldn't be able to come up with the answer - because we're not one of the WMs and couldn't possibly have that awareness. But there's sorta, kinda something about working with other people's point of view in there.

Bringsjord's bots actually did something rather different. They were apparently told that two of the three had been given a 'dumbing' pill that stopped them from being able to speak (actually a switch had been turned off; were the robots really clever enough to understand that distinction and the difference between a pill and a switch?); then they were asked 'did you get the dumbing pill?' Only one, of course, could answer, and duly answered 'I don't know': then, having heard its own voice, it was able to go on to say 'Oh, wait, now I know...!"

This test is obviously different from the original in many ways; it doesn't involve the same logic. Fairness, an essential factor in the original version, doesn't matter here, and in fact the test is egregiously unfair; only one bot can possibly win. The bot version seems to rest mainly on the robot being able to distinguish its own voice from those of the others (of course the others couldn't answer anyway; if they'd been really smart they would all have answered 'I wasn't dumbed', knowing that if they had been dumbed the incorrect conclusion would never be uttered). It does perhaps have a broadly similar sorta, kinda relation to awareness of points of view.

I don't propose to try to unpick the reasoning here any further: I doubt whether the experiment tells us much, but as presented in the New Scientist piece the logic is such a dog's breakfast and the details are so scanty it's impossible to get a proper idea of what is going on. I should say

that I have no doubt Ringsjord's actual presentation will be impeccably clear and well-justified in both its claims and its reasoning; foggy reports of clear research are more common than vice versa.

There's a general problem here about the slipperiness of defining human qualities. Ever since Plato attempted to define a man as 'a featherless biped' and was gleefully refuted by Diogenes with a plucked chicken, every definition of the special quality that defines the human mind seems to be torpedoed by counterexamples. Part of the problem is a curious bind whereby the task of definition requires you to give a specific test task; but it is the very non-specific open-ended generality of human thought you're trying to capture. This, I expect, is why so many specific tasks that once seemed definitively reserved for humans have eventually been performed by computers, which perhaps can do anything which is specified narrowly enough.

We don't know exactly what Bringsjord's bots did, and it matters. They could have been programmed explicitly just to do exactly what they did do, which is boring: they could have been given some general purpose module that does not terminate with the first answer and shows up well in these circumstances, which might well be of interest; or they could have been endowed with massive understanding of the real world significance of such matters as pills, switches, dumbness, wise men, and so on, which would be a miracle and raise the question of why Bringsjord was pissing about with such trivial experiments when he had such godlike machines to offer.

As I say, though, it's a general problem. In my view, the absence of any details about how the Room works is one of the fatal flaws in John Searle's Chinese Room thought experiment; arguably the same issue arises for the Turing Test. Would we award full personhood to a robot that could keep up a good conversation? I'm not sure I would unless I had a clear idea of how it worked.

I think there are two reasonable conclusions we can draw, both depressing. One is that we can't devise a good test for human qualities because we simply don't know what those qualities are, and we'll have to solve that imponderable riddle before we can get anywhere. The other possibility is that the specialness of human thought is permanently indefinable. Something about that specialness involves genuine originality, breaking the system, transcending the existing rules; so just as the robots will eventually conquer any specific test we set up, the human mind will always leap out of whatever parameters we set up for it.

But who knows, maybe the Ro-Man conference will surprise us with new grounds for optimism.

Stanford Consciousness

2 August 2015

The Stanford Encyclopaedia of Philosophy is twenty years old. It gives me surprisingly warm feelings towards Stanford that this excellent free resource exists. It's written by experts, continuously updated, and amazingly extensive. Long may it grow and flourish!

Writing an encyclopaedia is challenging, but an encyclopaedia of philosophy must take the biscuit. For a good encyclopaedia you need a robust analysis of the topics in the field so that they can be dealt with systematically, comprehensively, and proportionately. In philosophy there is never a consensus, even about how to frame the questions, never mind about what kind of answers might be useful. This must make it very difficult: do you try to cover the most popular schools of thought in an area? All the logically possible positions one might take up? A purely historical survey? Or summarise what the landscape is really like, inevitably importing your own preconceptions?

I've seen people complain that the SEP is not very accessible to newcomers, and I think the problem is partly that the subject is so protean. If you read an article in the SEP, you'll get a good view and some thought-provoking ideas; but what a noob looks for are a few pointers and landmarks. If I read a biography I want to know quickly about the subject's main works, their personal life, their situation in relation to other people in the field, the name of their theory or school, and so on. Most SEP subject articles cannot give you this kind of standard information in relation to philosophical problems. There is a real chance that if you read up a SEP article and then go and talk to professionals, they won't really get what you're talking about. They'll look at you blankly and then say something like:

"Oh, yes, I see where you're coming from, but you know, I don't really think of it that way…"

It's not because the article you read was bad, it's because everyone has a unique perspective on what the problem even is.

Let's look at Consciousness. The content page has:

- consciousness (Robert Van Gulick)
- animal (Colin Allen and Michael Trestman)
- higher-order theories (Peter Carruthers)
- and intentionality (Charles Siewert)
- representational theories of (William Lycan)
- seventeenth-century theories of (Larry M. Jorgensen)
- temporal (Barry Dainton)
- unity of (Andrew Brook and Paul Raymont)

All interesting articles, but clearly not a systematic treatment based on a prior analysis. It looks more like the set of articles that just happened to get written with consciousness as part of the subject. Animal consciousness, but no robot consciousness? Temporal consciousness, but no qualia or phenomenal consciousness? But I'm probably looking in the wrong place.

In Robert Van Gulick's main article we have something that looks much more like a decent shot at a comprehensive overview, but though he's done a good job it won't be a recognisable structure to anyone who hasn't read this specific article. I really like the neat division into descriptive, explanatory, and functional questions; it's quite helpful and illuminating: but you can't rely on anyone recognising it (Next time you meet a professor of philosophy ask him: if we divide the problems of consciousness into three, and the first two are descriptive and explanatory, what would the third be? Maybe he'll say 'Functional', but maybe he'll say 'Reductive' or something else - 'Intentional' or 'Experiential'; I'm pretty sure he'll need to think about it). Under 'Concepts of Consciousness' Van Gulick has 'Creature Consciousness': our noob would probably go away imagining that this is a well-known topic which can be mentioned in confident expectation of the implications being understood. Alas, no: I've read quite a few books about consciousness and can't immediately call to mind any other substantial reference to 'Creature Consciousness': I'm pretty sure that unless you went on to explain that you were differentiating it from 'State Consciousness' and 'Consciousness as an Entity', you might be misunderstood.

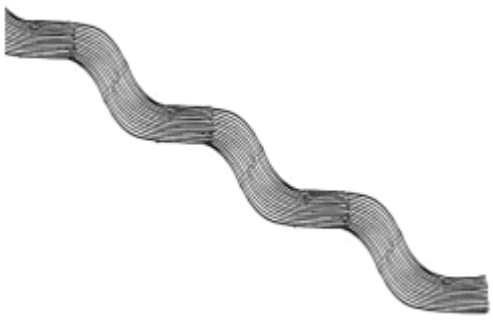
None of this is meant as a criticism of the piece: Van Gulick has done a great job on most counts (the one thing I would really fault is that the influence of AI in reviving the topic and promoting functionalist views is, I think, seriously underplayed). If you read the piece you will get about as good a view of the topic as that many words could give you, and if you're new to it you will run across some stimulating ideas (and some that will strike you as ridiculous). But when you next read a paper on philosophy of mind, you'll still have to work out from scratch how the problem is being interpreted. That's just the way it is.

Does that mean philosophy of mind never gets anywhere? No, I really don't think so, though it's outstandingly hard to provide proof of progress. In science we hope to boil down all the hypotheses to a single correct theory: in philosophy perhaps we have to be happy that we now have more answers (and more problems) than ever before.

And the SEP has got most of them! Happy Birthday!

Consciousness- A Stream?

8 August 2015



An interesting piece from Evan Thompson on the 'stream of consciousness'. The phrase is probably best known now as the name for a style of modern literary prose, but it originates with William James. Thompson compares James' concept of a smoothly rolling stream with the view taken by the Buddhist Abidharma tradition, which holds that closer consideration shows the stream to consist of discrete parts.

Thompson quotes two pieces of experimental evidence which broadly suggest the Abidharma view is closer to the truth. Experiments conducted by Francisco Varela on the young Thompson himself suggested that perception varied in harmony with the brain's alpha waves, although it seems the results have not been successfully replicated since. The other study related to the 'attentional blink' in which a stimulus rapidly following another is likely to be missed. It seems successful attempts by the subjects were accompanied by a kind of phase locking with theta rhythms; certain meditative techniques of mindfulness improved both the theta phase locking and the ability to perceive the following stimulus.

Overall, Thompson concludes that conscious perception isn't smoothly regular but comes in pulses. Perhaps we could say that it's more like the flow of a bloodstream than that of a river.

Still, though - is consciousness actually continuous? Suppose in fact that it was composed of a series of static moments, like the succeeding frames of a film. In a film the frames follow quickly, but we can imagine longer intervals if we like. However long the gaps, the story told by the film is unaffected and retains all its coherence; the discontinuity can only be seen by an observer outside the film. In the case of consciousness our experience actually is the succession of moments, so if consciousness were discontinuous we should never be aware of it directly. If we noticed anything at all, it would seem to us to be discontinuity in the external world.

It's not, of course, as simple as that; there are two particular issues. One is that consciousness is not automatically self-consciousness. To draw conclusions about our conscious state requires a second conscious state which is about the first one. We've remarked here before on Comte's objection that the second state necessarily disrupts the first, making reliable introspection impossible: James' view was that the second state had to be later, so that introspection was always retrospection.

This obviously raises many potential complications; all I want to do is pick out one possibility: that when we introspect the first and second order states alternate. Perhaps what we do is a moment of first-order thinking, then a moment of second order reflection on the moment just past, then another moment of simple first-order thought and so on; a process a bit like an artist flicking his gaze back and forth between subject and canvas.

If that's what happens, then it would clearly introduce a kind of pulse into our thoughts. This raises the curious possibility that our normal thoughts run smoothly but start to pulsate exactly when we start to think about them. The pulse would be an artefact of our own introspection.

The other issue is more fundamental. Both James and the Abidharma school apparently assume that our thoughts seem to come in a continuous flow. Well, mine don't. Yes, at times there is a coherent narrative sequence or a flowing perceptual experience, but these often seem like achievements of my concentration rather than the natural state of my mind. At least as often, things pop up unbidden, stop and start, and generally behave less like a flow and more like one damn thing after another. It's noteworthy that the stream of consciousness in literature is not characterised by smooth logical development, but by a succession of fragmentary ideas and perceptions, a more realistic picture in many ways.

However, reflecting on a train of thought afterwards we can sometimes see links that we didn't notice before. Several of our thoughts which seemed unrelated all bear on a particular anxiety or concern, say; scarcely a novel phenomenon in either psychology or literature. Hypothetically we might guess that our conscious moments are indeed part of a coherent stream, but one which includes important unconscious or subconscious elements, If we could see the whole process it might make fine logical sense, but all we get are the points where the undulating serpent's back breaks the surface.

Neither of those issues disturbs Thompson's modest conclusion that there is a kind of pulse on the surface of the stream; but there is deep water underneath, I think.

Inside Out

12 August 2015

The homunculus returns? I finally saw Inside Out (I seem to be talking about films a lot recently). Interestingly, it foregrounds a couple of problematic ways of thinking about the mind.

One, obviously, is the notorious homuncular fallacy. This is the tendency to explain mental faculties, say consciousness, by attributing them to a small entity within the mind - a "little man" that just has all the capacities of the whole human being. It's almost always condemned because it appears to do no more than defer the real explanation. If it's really a little man in your head that does consciousness, where does his consciousness come from? An even smaller man, in his head?

Inside Out of course does the homuncular thing very explicitly. The mind of the young girl Riley, the main character, where most of the action is set, is controlled by five primal emotions who are all fully featured cartoon people - Joy, Sadness, Anger, Fear, and Disgust, little people who walk around inside Riley's head doing the kind of thing people do (Is it actually inside her head? In the Beano's Numskulls cartoon, touted as a forerunner of Inside Out, much of the humour came from the definite physicality of the way they worked; here the five emotions view the world through a screen rather than eyeholes and use a console rather than levers. They could in fact be anywhere or in some undefined conceptual space.) It's an odd set (aren't Joy and Sadness the extremes of a spectrum?) Unexpectedly negative too: this is technically a Disney film, and it rates anger, fear, and disgust as more important and powerful than love? If it were full-on Disney the leading emotions would surely be Happy-go-lucky Feelin's, and Wishing on a Star.

There are some things to be said in favour of homunculi. Most people would agree that we contain a smaller entity that does all the thinking; the brain, or maybe even narrower than that (proponents of the Extended Mind would very much not agree, of course). Daniel Dennett has also spoken out for homunculi, suggesting that they're fine so long as the homunculi in each layer get simpler; in the end we get to ones that need no explanation. That's alright, except that I don't think the beings in this Dennettian analysis are really homunculi - they're more like black boxes. The true homunculus has all the capacities of a full human being rather than a simpler subset.

We see the problem that arises from that in Inside Out. The emotions are too rounded; they all seem to have a full set of feelings themselves; they show all show fear and Joy gets sad. How can that work?

The other thing that seems not quite right to me is unfortunately the climactic revelation that Sadness has a legitimate role. It is, apparently, to signal for help. In my view that can't really be the whole answer and the film unintentionally shows us the absurdity of the idea; it asks us to believe that being joyless, angry and withdrawn, behaving badly and running away are not enough to evoke concern and sympathetic attention from parents; you don't get your attention, and your hug till they see the tears.

No doubt sadness does often evoke support, but I can't think that's its main function. Funnily enough, Sadness herself briefly articulates a somewhat better idea early in the film. It's muttered so quickly I didn't quite get it, but it was something about providing an interval for adjustment and emotional recalibration. That sounds a bit more promising; I suspect it was

what a real psychologist told Pixar at some stage; something they felt they should mention for completeness but that didn't help the story.

Films and TV do shape our mental models; The Matrix laid down tramlines for many metaphysical discussions and Star Trek's transporters are often invoked in serious discussions of personal identity. Worse, fears about AI have surely been helped along by Hollywood's relentless and unimaginative use of the treacherous robot that turns on its creators. I hope Inside Out is not going to reintroduce homunculi to general thinking about the mind.

Closing The Gap

16 August 2015

A better neurophysiology, the answer to the Hard Problem? Kirchhoff and Hutto propose a slightly different way forward.

The Hard Problem, of course, is about reconciling the physical description of a conscious event with the way it feels from inside. This is the 'explanatory gap'. Most of us these days are monists of one kind or another; we believe the world ultimately consists of one kind of thing, usually matter, without a second realm of spirits or other metaphysical entities on top. Some people would, accordingly, seek to reduce the mental to the physical, perhaps even eliminating the mental so that our monism can be tidy (I'm a messy monist myself). Neurophysiology, as formulated by Varela and briefly described in Kirchhoff and Hutto's paper, does not look for a reduction, merely an explanation.

It does this by putting aside any idea of representations or computations; instead, it proposes a practical research programme in which introspective reports of experience are matched with scans or other physical investigations. By elucidating the structure of both experience and physical event, the project aims to show how the two sides of experience constrain each other.

This, though, doesn't seem enough for Kirchhoff and Hutto. Researching the two sides of the matter together is fine, but how will it ever show constraints, or generate an explanation? it seems it will be doomed to merely exhibiting correlation. Moreover, rather than resolving the explanatory gap, this approach seems to consolidate it.

These are reasonable objections, but I don't think it's quite as hopeless as that. The aspiration must surely be that the exploration comes together by exhibiting, not just correlation, but an underlying identity of structure? We might hope that the physical structure of the visual cortex tells us something about our colour space and visual experience that matches the structure of 0ur direct experience of colour, for example, in such a way that the mysterious quality of that experience is attenuated and eventually even dispelled. Other kinds of explanation might emerge. When I take off my glasses and look at the surface of brightly lit swimming pool, I see a host of white circles, all the same size and filled with the suggestion of a moire pattern, bobbing daintily about. In a pre-scientific era, this would have been hard to account for, but now I know it is entirely the result of some facts about the shape of my eyes and the lenses in them, and phenomenological worries don't even get started. It could be that neurophilosophy can succeed in offering explanations good enough to remove the worries that currently exist. The great thing about it, of course, is that even if that hope is philosophically misplaced, elucidating the structure of experience from both ends is a very worthwhile project anyway, one that can surely only yield valuable new understanding.

However, what Kirchhoff and Hutto propose is that we go a little further and abolish the gap. Instead of affirming the separateness of the physical and the phenomenal, they suggest, we should recognise that they represent to different descriptions of a single thing.

That might seem a modest adjustment, but they also assert that the phenomenal character of experience actually arises not from the mere physics, but from the situation of that experience, taking place in an enactive, embodied context. So if we hold a book, we can see it; if we shut our

eyes, we continue to feel it; but we also have a more complex engagement with it from our efforts to hold up what we know is a book, the feel of pages, and so on. There's all sorts of stuff going on that isn't the mere physical contact, and that's what yields the character of the experience.

I see that, I think, but it's a little odd. If we imagine floating in a sensory deprivation tank and gazing at a smooth, uniform red wall, we seem to be free of a lot of the context we'd normally have and on this view it's a bit hard to see where the phenomenal intensity would be coming from (perhaps from the remembered significance of red?) We might suspect that Kirchhoff and Hutto are getting their phenomenal content smuggled in with the more complex phenomenal experience that they implicitly demand by requiring context, an illicit supplement that remains unexplained.

On this, why not let a thousand flowers grow; go ahead and develop explanations according to any exploratory project you prefer, and then we'll have a look. Some of them might be good even if your underlying theory is wrong.

I think it is, incidentally. For me the explanatory gap is always misconstrued; the real gap is not between physics and phenomenology, it's between theory and actuality, something that shouldn't puzzle us, or at least not in the way it always does.

Scott's Aliens

23 August 2015

Scott Bakker has given an interesting new approach to his Blind Brain Theory (BBT): in two posts on his blog he considers what kind of consciousness aliens could have, and concludes that the process of evolution would put them into the same hole where, on his view, we find ourselves.

BBT, in sketchy summary, says that we have only a starvation diet of information about the cornucopia that really surrounds us; but the limitations of our sources and cognitive equipment mean we never realise it. To us it looks as if we're fully informed, and the glitches of the limited heuristics we use to cobble together a picture of the world, when turned on ourselves in particular, look to us like real features. Our mental equipment was never designed for self-examination and attempting metacognition with it generates monsters; our sense of personhood, agency, and much about our consciousness comes from the deficits in our informational resources and processes.

Scott begins his first post by explaining his own journey from belief in intentionalism to eliminativist scepticism about it, and sternly admonishes those of us still labouring in intentionalist error for our failure to produce a positive account of how human minds could have real intentionality.

What about aliens - Scott calls the alien players in his drama 'Thespians' - could they be any better off than we are? Evolution would have equipped them with senses designed to identify food items, predators, mates, and so on; there would be no reason for them to have mental or sensory modules designed to understand the motion of planets or stars, and turning their senses on their own planet would surely tell them incorrectly that it was motionless. Scott points out that Aristotle's argument against the movement of the Earth is rather good: if the Earth were moving, we should see shifts in the relative position of the stars, just as the relative position of objects in a landscape shifts when we we view them from the window of a moving train; yet the stars remain precisely fixed. The reasoning is sound; Aristotle simply did not know and could not imagine the mind-numbingly vast distances that make the effect invisibly small to unaided observation. The unrealised lack of information led Aristotle into misapprehension, and it would surely do the same for the Thespians; a nice warm-up for the main argument.

Now it's a reasonable assumption that the Thespians would be social animals, and they would need to be able to understand each other. They'd get good at what is often somewhat misleadingly called theory of mind; they'd attribute motives and so on to each other and read each others behaviour in a fair bit of depth. Of course they would have no direct access to other Thespians; actual inner workings. What happens when they turn their capacity for understanding other people on themselves? In Scott's view, plausibly enough, they end up with quite a good practical understanding whose origins are completely obscure to them; the lashed-up mechanisms that supply the understanding neither available to conscious examination or in fact even visible.

This is likely enough, and in fact doesn't even require us to think of higher cognitive faculties. How do we track a ball flying through the air so we can catch it? Most people would be hard put to describe what the brain does to achieve that, though in practice we do it quite well. In fact,

those who could put down the algorithm most likely get it wrong too, because it turns out the brain doesn't use the optimal method: it uses a quick and easy one that works OK in practice but doesn't get your hand to the right place as quickly as it could.

For Scott all this leads to a gloomy conclusion; much of our view about what we are and our mental capacities is really attributable to systematic error, even to something we could regard as a cognitive deficit or disease. He cogently suggests how dualism and other errors might arise from our situation.

I think the Thespian account is the most accessible and persuasive account Scott has given to date of his view, and it perhaps allows us to situate it better than before. I think the scope of the disaster is a little less than Scott supposes, in two ways. First, he doesn't deny that routine intentionality actually works at a practical level, and I think he would agree we can even hope to give a working level description of how that goes. My own view is that it's all a grand extension of our capacity for recognition, (and I was more encouraged than not by my recent friendly disagreement with Marcus Perlman over on Aeon Ideas; I think his use of the term 'iconic' is potentially misleading, but in essence I think the views he describes are right and enlightening) but people here have heard all that many times. Whether I'm right or not we probably agree that some practical account of how the human mind gets its work done is possible.

Second, on a higher level, it's not completely hopeless. We are indeed prone to dreadful errors and to illusions about how our minds work that cannot easily be banished. But we don't have to resort to those flawed heuristics; e can take our pure basic understanding and try again, either through some higher meta-meta-cognition or by simply standing aside and looking at the thing from a scientific third-person point of view. Aristotle was wrong, but we got it right in the end, and why shouldn't Scott's own view be part of getting it righter about the mind? I don't think he would disagree on that either (he'll probably tell us); but he feels his conclusions have disastrous implications for our sense of what we are.

Here we strike something that came up in our recent discussion of free will and the difference between determinists and compatibilists. It may be more a difference of temperament than belief. people like me say, OK, no, we don't have the magic abilities we looked to have, so let's give those terms a more sensible interpretation and go merrily on our way. The determinists, the eliminativists, agree that the magic has gone - in fact they insist - but they sit down by the roadside, throw ashes on their heads, and mourn it. They share with the naive, the libertarians, and the believers in a magic power of intentionality, the idea that something essential and basically human is lost when we move on in this way. Perhaps people like me came in to have the magic explained and are happy to see the conjuring tricks set out; others wanted the magic explained and *for it to remain magic*?

Freedom – Why Worry?

27 August 2015

Why does the question of determinism versus free will continue to trouble us? There's nothing too strange, perhaps, about a philosophical problem remaining in play for a while - or even for a few hundred years: but why does this one have such legs and still provoke such strong and contrary feelings on either side?

For me the problem itself is solved – and the right solution, broadly speaking, has been known for centuries: determinism is true, but we also have free choice in a meaningful sense. St Augustine, to go no earlier, understood that free will and predestination are not contradictory, but people still find it confusing that he spoke up for both.

If this view - compatibilism - is right, why hasn't it triumphed? I'm coming to think that the strongest opposition on the question might not in fact be between the hard-line determinists and the uncompromising libertarians but rather a matter of both ends against the middle. Compatibilists like me are happy to see the problem solved and determinism reconciled with common sense, whereas people from both the extremes insist that that misses something crucial. They believe the 'loss' of free will radically undercuts and changes our traditional sense of what we are as human beings. They think determinism, for better or worse, wipes away some sacred mark from our brows. Why do they think that?

Let's start by quickly solving the old problem. Part one: determinism is true. It looks, with some small reservations about the interpretation of some esoteric matters, as if the laws of physics completely determine what happens. Actually even if contemporary physics did not seem to offer the theoretical possibility of full determination, we should be inclined to think that some set of rules did. A completely random or indeterminate world would seem scarcely distinguishable from a nullity; nothing definite could be said about it and no reliable predictions could be made because everything could be otherwise. That kind of scenario, of disastrous universal incoherence is extreme, and I admit I know of no absolute reason to rule out a more limited, demarcated indeterminacy. Still, the idea of delimited patches of randomness seems odd, inelegant and possibly problematic. God, said Einstein, does not play dice.

Beyond that, moreover, there's a different kind of point. We came into this business in pursuit of truth and knowledge, so it's fair to say that if there seemed to be patches of uncertainty we should want to do our level best to clarify them out of existence. In this sense it's legitimate to regard determinism not just as a neutral belief, but as a proper aspiration. Even if we believe in free will, aren't we going to want a theory that explains how it works, and isn't that in the end going to give us rules that determine the process? Alright, I'm coming to the conclusion too soon: but in this light I see determinism as a thing that lovers of truth must strive towards (even if in vain) and we can note in passing that that might be some part of the reason why people champion it with zeal.

We're not done with the defence, anyway. One more thing we can do against indeterminacy is to invoke the deep old principle which holds that nothing comes of nothing, and that nothing

therefore happens unless it must; if something particular must happen, then the compulsion is surely encapsulated in some laws of nature.

Further still, even if none of that were reliable, we could fall back on a fatalistic argument. If it is true that on Tuesday you'll turn right, then it was true on Monday that you would turn right on Tuesday; so your turning that way rather than left was already determined.

Finally, we must always remember that failure to establish determinism is not success in establishing liberty. Determinism looks to be true; we should try to establish its truth if by any means we can: but even if we fail, that failure in itself leaves us not with free will but with an abhorrent void of the unknowable.

Part two: we do actually make free decisions. Determinism is true, but it bites firmly only at a low level of description; not truly above the level of particles and forces. To look for decisions or choices at that level is simply a mistake, of the same general kind as looking for bicycles down there. Their absence from the micro level does not mean that cyclists are systematically deluded. Decisions are processes of large neural structures, and I suggest that when we describe them as free we simply mean the result was not constrained externally. If I had a gun to my head or my hands were tied, then turning left was not a free decision. If no-one could tell which way I should go without knowledge of what was going on in the large neural structures that realise my mind, then it was free. There are of course degrees of freedom and plenty of grey areas, but the essential idea is clear enough. Freedom is just the absence of external constraint on a level of description where people and decisions are salient, useful concepts.

For me, and I suppose other compatibilists, that's a satisfying solution and matches well with what I think I've always meant when I talk about freedom. Indeed, it's hard for me to see what else freedom could mean. What if God did play dice after all? Libertarians don't want their free decisions to be random, they want them to belong to them personally and reflect consideration of the circumstances; the problem is that it's challenging for them to explain in that case how the decisions can escape some kind of determination. What unites the libertarians and the determinists is the conviction that it's that inexplicable, paradoxical factor we are concerned to affirm or deny, and that its presence or absence says something important about human nature. To eliminate the magic, as compatibilists do, is on their view to shoot the fox and spoil the hunt. What are they both so worried about?

I speculate that the one factor here is a persistent background confusion. Determinism, we should remember, is an intellectual achievement, both historically and often personally. We live in a world where nothing much about human beings is certainly determined; only careful reflection reveals that in the end, at the lowest level of detail and at the very last knockings of things, there must be certainty. This must remain a theoretical conclusion, certainly so far as human beings are concerned; our behaviour may be determinate, but it is not determinable; certainly not in practice and very probably not even in theory, given the vast complexity, chaotic organisation and marvellously emergent properties of our brains. Some of those who deny determinism may be moved, not so much by explicit rejection of the true last-ditch thesis, but by the certainty that our minds are not determinable by us or by anyone. This muddying of the waters is perpetuated even now by arguments about how our minds may be strongly influenced by high-level factors: peer pressure, subliminal advertising, what we were given to read just before making a decision. These arguments may be presented in favour of determinism together with the water-tight last-ditch case, but they are quite different, and the high-level determinism they support is not certainly true but rather an eminently deniable hypothesis. In the end our

behaviour is determined, but can we be programmed like robots by higher level influences? Maybe in some cases - generally, probably not.

The second, related factor is a certain convert enthusiasm. If determinism is a personal intellectual achievement it may well be that we become personally invested in it. When we come to appreciate its truth for the first time it may seem that we have grasped a new perspective and moved out of the confused herd to join the scientifically enlightened. I certainly felt this on my first acquaintance with the idea; I remember haranguing a friend about the truth of determinism in a way that must, with hindsight, have resembled religious conviction and been very tiresome.

“Yes, yes, OK, I get it,” he would say in a vain attempt to stop the flow.

Now no-one lives pure determinism; we all go behaving as if agency and freedom were meaningful. The fact that this involves an unresolved tension between your philosophy and the ideas about people you actually live by was not a deterrent to me then, however; in fact it may even have added a glamorous sheen of esoteric heterodoxy to the whole thing. I expect other enthusiasts may feel the same today. The gradual revelation, some years later, that determinism is true but actually not at all as important as you thought is less exciting: it has rather a dying fall to it and may be more difficult to assimilate. Consistency with common sense is perhaps a game for the middle aged.

“You know, I’ve been sort of nuancing my thinking about determinism lately…”

“Oh, what, Peter? You made me live through the conversion experience with you - now I have to work through your apostasy, too?”

On the libertarian side, it must be admitted that our power of decision really does look sort of strange, with a power far exceeding that of mere absence of constraint. There are at least two reasons for this. One is our ability to use intentionality to think about anything whatever, and base our decisions on those thoughts. I can think about things that are remote, non-existent, or even absurd, without any difficulty. Most notably, when I make decisions I am typically thinking about future events: will I turn left or right tomorrow? How can future events influence my behaviour now?

It’s a bit like the time machine case where I take the text of Hamlet back in time and give it to Shakespeare - who never actually produced it but now copies it down and has it performed. Who actually wrote it, in these circumstances? No-one, it just appeared at some point. Our ability to think about the future, and so use future goals as causes of actions now, seems in the same way to bring our decisions into being out of nowhere inside us. There was no prior cause, only later ones, so it really seems as if the process inverts and disrupts the usual order of causality.

We know this is indeed remarkable but it isn’t really magic. On my view it’s simply that our recognition of various entities that extend over time allows a kind of extrapolation. The actual causal processes, down at that lowest level, tick away in the right order, but our pattern-matching capacity provides processes at a higher level which can legitimately be said to address the future without actually being caused by it. Still, the appearance is powerful, and we may be impatient with the kind of materialist who prefers to live in a world with low ceilings, insists on everything being material and denies any independent validity to higher levels of description. Some who think that way even have difficulty accepting that we can think directly

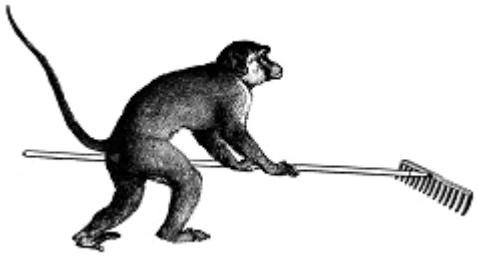
about mathematical abstractions - quite a difficult posture for anyone who accepts the physics that draws heavily on them.

The other thing is the apparent, direct reality of our decisions. We just know we exercise free will, because we experience the process immediately. We can feel ourselves deciding. We could be wrong about all sorts of things in the world, but how could I be wrong about what I think? I believe the feeling of something ineffable here comes from the fact that we are not used to dealing with reality. Most of what we know about the world is a matter of conscious or unconscious inference, and when we start thinking scientifically or philosophically it is heavily informed by theory. For many people it starts to look as if theory is the ultimate bedrock of things, rather than the thin layer of explanation we place on top. For such a mindset the direct experience of one's own real thoughts looks spooky; its particularity, its haecceity, cannot be accounted for by theory and so looks anomalous. There are deep issues here, but really we ought not to be foxed by simple reality.

That's it, I think: that really is it done and dusted. More could be said, of course, more will be said. The issues above are like optical illusions: just knowing how they work doesn't make them go away, and so minds will go on boggling. People will go on furiously debating free will: that much is both determined and determinable.

Do We Care About Where?

31 August 2015



Do we care whether the mind is extended? The latest issue of the JCS features papers on various aspects of extended and embodied consciousness.

In some respects I think the position is indicated well in a paper by Michael Wheeler, which tackles the question of whether phenomenal experience is realised, at least in part, outside the brain. One reason I think this attempt is representative is its huge ambition. The general thesis of extension is that it makes sense to regard tools and other bodily extensions - the iPad in my hand now, but also simple notepads, and even sticks - as genuine participating contributors to mental events. This is sort of appealing if we're talking about things like memory, or calculation, because recording data and doing sums are the kind of thing the iPad does. Even for sensory experience it's not hard to see how the camera and Skype might reasonably be seen as extended parts of my perceptual apparatus. But phenomenal experience, the actual business of how something feels? Wheeler notes a strong intuition that this, at least, must be internal (here read as 'neural'), and this surely comes from the feeling that while the activity of a stick or pad looks like the sort of thing that might be relevant to "easy problem" cognition, it's hard to see what it could contribute to inner experience. Granted, as Wheeler says, we don't really have any clear idea what the brain is contributing either, so the intuition isn't necessarily reliable. Nevertheless it seems clear that tackling phenomenal consciousness is particularly ambitious, and well calculated to put the overall idea of extension under stress.

Wheeler actually examines two arguments, both based on experiments. The first, from Noë, relies on sensory substitution. Blind people fitted with apparatus that delivers optical data in tactile form begin to have what seems like visual experience (How do we know they really do? Plenty of scope for argument, but we'll let that pass.) The argument is that the external apparatus has therefore transformed their phenomenal experience.

Now of course it's uncontroversial that changing what's around you changed the content of your experience, and changing the content changes your phenomenal experience. The claim here is that the whole modality has been transformed, and without a parallel transformation in the brain. It's the last point that seems particularly vulnerable. Apparently the subjects adapt quickly to the new kit, too quickly for substantial 'neural rewiring', but what's substantial in this context? There are always going to be some neural changes during any experience, and who's to say that those weren't the crucial ones?

The second argument is due to Kiverstein and Farina, who report that when macaques are trained to use rakes to retrieve food, the rakes are incorporated into their body image (as reflected in neural activity). This is easy enough to identify with - if you use a stick to probe the ground, you quickly start to experience the 'feel' of the ground as being at the end of the stick, not in your hand. Does it prove that your phenomenal experience is situated partly in the stick? Only in a sense that isn't really the one required - we already felt it as being in the hand. We never experience tactile sensation as being in the brain: the anti-extension argument is merely that the brain is uniquely the substrate where it the feeling is generated.

Wheeler, rather anti-climatically but I think correctly, thinks neither argument is successful; and that's another respect in which I think his paper represents the state of the extended mind thesis; both ambitious and unproven.

Worse than that, though, it illustrates the point which kills things for me; I don't really care one way or the other. Shall we call these non-neural processes mental? What if we do? Will we thereby get a new insight into how mental processes work? Not really, so why worry? The thesis that experience is external in a deeper sense, external to my mind, is strange and mind-boggling; the thesis that it's external in the flatly literal sense of having some of its works outside the brain is just not that philosophically interesting.

OK, it's true that what we know about the brain doesn't seem to explicate phenomenal experience either, and perhaps doesn't even look like the kind of thing that in principle might do so. But if there are ever going to be physical clues, that's kind of where you'd bet on them being.

Is phenomenal experience extended? Well, I reckon phenomenal experience is tied to the non-phenomenal variety. Red qualia come with the objective perception of red. So if we accept the extended mind for the latter, we should probably accept it for the former. But please yourself; in the absence of any additional illumination, who cares where it is?

Selves Without Concepts

6 September 2015

Thinking without concepts is a strange idea at first sight; isn't it always the concept of a thing that's involved in thought, rather than the thing itself? But I don't think it's really as strange as it seems. Without looking too closely into the foundations at this stage, I interpret conceptual thought as simply being one level higher in abstraction than its non-conceptual counterpart.

Dogs, I think, are perfectly capable of non-conceptual thinking. Show them the lead or rattle the dinner bowl and they assent enthusiastically to the concrete proposal. Without concepts, though, dogs are tied to the moment and the actual; it's no good asking the dog whether it would prefer walkies tomorrow morning or tomorrow afternoon; the concept cannot gain a foothold - nothing more abstract than walkies now can really do so. That doesn't mean we should deny a dog consciousness - the difference between a conscious and unconscious dog is pretty clear - only to certain human levels. The advanced use of language and symbols certainly requires concepts, though I think it is not synonymous with it and conceptual but inexplicit thought seems a viable option to me. Some, though, have thought that it takes language to build meaningful self-consciousness.

Kristina Musholt has been looking briefly at whether self-consciousness can be built out of purely non-conceptual thought, particularly in response to Bermudez, and summarising the case made in her book Thinking About Oneself.

Musholt suggests that non-conceptual thought reflects knowledge-how rather than knowledge-that; without quite agreeing completely about that we can agree that non-conceptual thought can only be expressed through action and so is inherently about interaction with the world, which I take to be her main pointer.

Following Bermudez it seems we are to look for three things from any candidate for self-awareness; namely,

(1) non-accidental self-reference,

(2) immediate action relevance, and

(3) immunity to error through misidentification.

That last one may look a bit scary; it's simply the point that you can't be wrong about the identity of the person thinking your own thoughts. I think there are some senses in which this ain't necessarily so, but for present purposes it doesn't really matter. Bermudez was concerned to refute those who think that self-consciousness requires language; he thought any such argument collapses into circularity; to construct self-consciousness out of language you have to be able to talk about yourself, but talking about yourself requires the very self-awareness you were supposed to be building.

Bermudez, it seems, believes we can go elsewhere and get our self-awareness out of something like that implicit certainty we mentioned earlier. As thought implies the thinker, non-conceptual thoughts will serve us perfectly well for these purposes. Musholt, though broadly in sympathy, isn't happy with that. While the self may be implied simply by the existence of non-conceptual thoughts, she points out that it isn't represented, and that's what we really need. For one thing, it makes no sense to her to talk about immunity from error when it applies to something that isn't even represented - it's not that error is possible, it's that the whole concept of error or immunity doesn't even arise.

She still wants to build self-awareness out of non-conceptual thought, but her preferred route is social. As we noted she thinks non-conceptual thought is all about interaction with the world, and she suggests that it's interaction with other people that provides the foundation for our awareness ourselves. It's our experience of other people that ultimately grounds our awareness of ourselves as people.

That all seems pretty sensible. I find myself wondering about dogs, and about the state of mind of someone who grew up entirely alone, never meeting any other thinking being. It's hard even to form a plausible thought experiment about that, I think. The human mind being what it is, I suspect that if no other humans or animals were around inanimate objects would be assigned imaginary personalities and some kind of substitute society cobbled up. Would the human being involved end up with no self-awareness, some strangely defective self-awareness (perhaps subject to some kind of dissociative disorder?), or broadly normal? I don't even have any clear intuition on the matter.

Anyway, we should keep track of the original project, which essentially remains the one set out by Bermudez. Even if we don't like Musholt's proposal better than his, it all serves to show that there is actually quite a rich range of theoretical possibilities here, which tends to undermine the view that linguistic ability is essential. To me it just doesn't seem very plausible that language should come before self-awareness, although I think it does come before certain forms of self-awareness. The real take-away, perhaps, is that self-awareness is a many-splendoured thing and different forms of it exist on all the three levels mentioned and surely more, too. This conclusion partly vindicates the attack on language as the only basis for self-awareness, undercutting Bermudez's case for circularity. If self-awareness actually comes in lots of forms, then the sophisticated, explicit, language-based kind doesn't need to pull itself up by its bootstraps, it can grow out of less explicit versions.

Anyway, Musholt has at least added to our repertoire a version of self-awareness which seems real and interesting - if not necessarily the sole or most fundamental version.

Ethics In A Nutshell

13 September 2015

We've done so much here towards clearing up the problems of consciousness I thought we might take a short excursion and quickly sort out ethics?

It's often thought that philosophical ethics has made little progress since ancient times; that no firm conclusions have been established and that most of the old schools, along with a few new ones, are still at perpetual, irreconcilable war. There is some truth in that, but I think substantial progress has been made. If we stop regarding the classic insights of different philosophers as rivals and bring them together in a synthesis, I reckon we can put together a general ethical framework that makes a great deal of sense.

What follows is a brief attempt to set out such a synthesis from first principles, in simple non-technical terms. I'd welcome views: it's only fair to say that the philosophers whose ideas I have nicked and misrepresented would most surely hate it.

The deepest questions of philosophy are refreshingly simple. What is there? How do I know? And what should I do?

We might be tempted to think that that last question, the root question of ethics, could quickly be answered by another simple question; what do you want to do? For thoughtful people, though, that has never been enough. We know that some of the things people want to do are good, and some are bad. We know we should avoid evil deeds and try to do good ones - but it's sometimes truly hard to tell which they are. We may stand willing to obey the moral law but be left in real doubt about its terms and what it requires. Yet, coming back to our starting point, surely there really is a difference between what we want to do and what we ought to do?

Kant thought so: he drew a distinction between categorical and hypothetical imperatives. For the hypothetical ones, you have to start with what you want. If you're thirsty, then you should drink. If you want to go somewhere, then you should get in your car. These imperatives are not ethical; they're simply about getting what you want. The categorical imperative, by contrast, sets out what you should do anyway, in any circumstances, regardless of what you want; and that, according to Kant, is the real root of morality.

Is there anything like that? Is there anything we should unconditionally do, regardless of our aims or wishes? Perhaps we could say that we should always do good; but even before we get on to the difficult task of defining 'good', isn't that really a hypothetical imperative? It looks as if it goes: if you want to be good, behave like this...? Why do we have to be good? Let's imagine that Kant, or some other great man, has explained the moral law to us so well, and told us what good is, so correctly and clearly that we understand it perfectly. What's to stop us exercising our freedom of choice and saying "I recognise what is good, and I choose evil"?

To choose evil so radically and completely may sound more like a posture than a sincere decision - too grandly operatic, too diabolical to be altogether convincing - but there are other, appealing ways we might want to rebel against the whole idea of comprehensive morality. We might just seek some flexibility, rejecting the idea that morality rules our lives so completely, always telling us exactly what to do at every turn. We might go further and claim unrestricted freedom, or we might think that we may do whatever we like so long as we avoid harm to others, or do not commit actual crimes. Or we might decide that morality is simply a matter of useful social convention, which we propose to go along with just so long as it suits our chosen path, and no further. We might come to think that a mature perspective accepts that we don't need to be perfect; that the odd evil deed here and there may actually enhance our lives and make us more rounded, considerable and interesting people.

Not so fast, says Kant, waving a finger good-naturedly; you're missing the point; we haven't yet even considered the nature of the categorical imperative! It tells us that we must act according to the rules we should be happy to see others adopt. We must accept for ourselves the rules of behaviour we demand of the people around us.

But why? It can be argued that some kind of consistency requires it, but who said we had to be consistent? Equally, we might well argue that fairness requires it, but we haven't yet been given a reason to be fair, either. Who said that we had to act according to any rule? Or even if we accept that, we might agree that everyone should behave according to rules we have cunningly slanted in our own favour (Don't steal, unless you happen to be in the special circumstances where I find myself to be) or completely vacuous rules (Do whatever you want to do). We still seem to have a serious underlying difficulty: why be good? Another simple question, but it's one we can't answer properly yet.

For now, let's just assume there is something we ought to do. Let's also assume it is something general, rather than a particular act on a particular occasion. If the single thing we ought to do were to go up the Eiffel Tower at least once in our life, our morality would be strangely limited and centred. The thing we ought to do, let's assume, is something we can go on doing, something we can always do more of. To serve its purpose it must be the kind of behaviour that racks up something we can never have too much of.

There are people who have ethical theories which are exactly based on general goals like that, namely consequentialists. They believe the goodness of our acts depends on their consequences. The idea is that our actions should be chosen so that as a consequence some general desideratum is maximised. The desired thing can vary but the most famous example is the happiness which Jeremy Bentham embodied in the Utilitarians' principle: act so as to bring about the greatest happiness of the greatest number of people.

Old-fashioned happiness Utilitarianism is a simple and attractive theory, but there are several problems with the odd results it seems to produce in unusual cases. Putting everyone in some kind of high-tech storage ward but constantly stimulating the pleasure centres in their brains with electrodes appears a very good thing indeed if we're simply maximising happiness. All those people spend their existence in a kind of blissful paralysis: the theory tells us this is an excellent result, something we must strive to achieve, but it surely isn't. Some kinds of ecstatic madness, indeed, would be high moral achievements according to simple Utilitarianism.

Less dramatically, people with strong desires, who get more happiness out of getting what they want, are awarded a bigger share of what's available under utilitarian principles. In the extreme

case the needs of ‘happiness monsters’ whose emotional response is far greater than anyone else’s, come to dominate society. This seems strange and unjust; but perhaps not to everyone. Bentham frowns at the way we’re going: why, he asks, should people who don’t care get the same share as those who do?

That case can be argued, but it seems the theory now wants to tutor and reshape our moral intuitions, rather than explaining them. It seems a real problem, as some later utilitarians recognised, that the theory provides no way at all of judging one source or kind of happiness better or worse than another. Surely this reduces and simplifies too much; we may suspect in fact that the theory misjudges and caricatures human nature. The point of life is not that people want happiness; it’s more that they usually get happiness from having the things they actually want.

With that in mind, let’s not give up on utilitarianism; perhaps it’s just that happiness isn’t quite the right target? What if, instead, we seek to maximise the getting of what you want - the satisfaction of desires? Then we might be aiming a little more accurately at the real desideratum, and putting everyone in pleasure boxes would no longer seem to be a moral imperative; instead of giving everyone sweet dreams, we have to fulfil the reality of their aspirations as far as we can.

That might deal with some of our problems, but there’s a serious practical difficulty with utilitarianism of all kinds; the impossibility of knowing clearly what the ultimate consequences of any action will be. To feed a starving child seems to be a clear good deed; yet it is possible that by evil chance the saved infant will grow up to be a savage dictator who will destroy the world. If that happens the consequences of my generosity will turn out to be appalling. Even if the case is not as stark as that, the consequences of saving a child roll on through further generations, perhaps forever. The jury will always be out, and we’ll never know for sure whether we really brought more satisfaction into the world or not.

Those are drastic cases, but even in more everyday situations it’s hard to see how we can put a numerical value on the satisfaction of a particular desire, or find any clear way of rating it against the satisfaction of a different one. We simply don’t have any objective or rigorous way of coming up with the judgements which utilitarianism nevertheless requires us to make.

In practice, we don’t try to make more than a rough estimate of the consequences of our actions. We take account of the obvious immediate consequences: beyond that the best we can do is to try to do the kind of thing that in general is likely to have good consequences. Saving children is clearly good in the short term, and people on the whole are more good than bad (certainly for a utilitarian - each person adds more satisfiable desires to the world), so that in most cases we can justify the small risk of ultimate disaster following on from saving a life.

Moreover, even if I can’t be sure of having done good, it seems I can at least be sure of having acted well; I can’t guarantee good deeds but I can at least guarantee being a good person. The goodness of my acts depend on their real consequences; my own personal goodness depends only on what I intended or expected, whether things actually work out the way I thought they would or not. So if I do my best to maximise satisfaction I can at least be a good person, even if I may on rare occasions be a good person who has accidentally done bad things.

Now though, if I start to guide my actions according to the kind of behaviour that is likely to bring good results, I am in essence going to adopt rules, because I am no longer thinking about individual acts, but about general kinds of behaviour. Save the children; don’t kill; don’t steal.

Utilitarianism of some kind still authorises the rules, but I no longer really behave like a Utilitarian; instead I follow a kind of moral code.

At this point some traditional-looking folk step forward with a smile. They have always understood that morality was a set of rules, they explain, and proceed to offer us the moral codes they follow, sanctified by tradition or indeed by God. Unfortunately on examination the codes, although there are striking broad resemblances, prove to be significantly different both in tone and detail. Most of them also seem, perhaps inevitably, to suffer from gaps, rules that seem arbitrary, and ones which seem problematic in various other ways.

How are we to tell what the correct code is? Our code is to be authorised and judged on the basis of our preferred kind of utilitarianism, so we will choose the rules that tend to promote the goal we adopted provisionally; the objective of maximising the satisfaction of desires. Now, in order to achieve the maximal satisfaction of desires, we need as many people as possible living in comfortable conditions with good opportunities and that in turn requires an orderly and efficient society with a prosperous economy. We will therefore want a moral code that promotes stable prosperity. There turns out to be some truth in the suggestion that morality in the end consists of the rules that suit social ends! Many of these rules can be worked out more or less from first principles. Without consistent rules of property ownership, without reasonable security on the streets, we won't get a prosperous society and economy, and this is a major reason why the codes of many cultures have a lot in common.

There are also, however, many legitimate reasons why codes are different, too. In certain areas the best rules are genuinely debatable. In some cases, moreover, there is genuinely no fundamental reason to prefer one reasonable rule over another. In these cases it is important that there are rules, but not important what they are - just as for traffic regulations it is not important whether the rule is to drive on the left or the right, but very important that it is one or the other. In addition the choice of rules for our code embodies some assumptions about human nature and behaviour and which arrangements work best with it. Ethical rules about sexual behaviour are often of this kind, for example. Tradition and culture may have a genuine weight in these areas, another potentially legitimate reason for variation in codes.

We can also make a case for one-off exceptions. If we believed our code was the absolute statement of right and wrong, perhaps even handed down by God, we should have no reason to go against it under any circumstances. Anything we did that didn't conform with the code would automatically be bad. We don't believe that, though; we've adopted our code only as a practical response to difficulties with working out what's right from first principles - the impossibility of determining what the final consequences of anything we do will be. In some circumstances, that difficulty may not be so great. In some circumstances it may seem very clear what the main consequences of an action will be, and if it looks more or less certain that following the code will, in a particular exceptional case, have bad consequences, we are surely right to disobey the code; to tell white lies, for example, or otherwise bend the rules. This kind of thing is common enough in real life, and I think we often feel guilty about it. In fact we can be reassured that although the judgements required may sometimes be difficult, breaking the code to achieve a good result is the right thing to do.

The champions of moral codes find that hard to accept. In their favour we must accept that observance of the code generally has a significant positive value in itself. We believe that following the rules will generally produce the best results; it follows that if we set a bad example or undermine the rules by flouting them we may encourage disobedience by others (or just lax

habits in ourselves) and so contribute to bad outcomes later on. We should therefore attach real value to the code and uphold it in all but exceptional cases.

Having got that far on the basis of a provisional utilitarianism, we can now look back and ask whether the target we chose, that of maximising the satisfaction of desires, was the right one. We noticed that odd consequences follow if we seek to maximise happiness; direct stimulation of the pleasure centres looks better than living your life, happiness monsters can have desires so great that they overwhelm everything else. It looked as if these problems arise mainly in situations where the pursuit of simple happiness is too narrowly focused, over-riding other objectives which also seem important.

In this connection it is productive to consider what follows if we pursue some radical alternative to happiness. What, indeed, if we seek to maximise misery? The specifics of our behaviour in particular circumstances will change, but the code authorised by the pursuit of unhappiness actually turns out to be quite similar to the one produced by its opposite. For maximum misery, we still need the maximum number of people. For the maximum number of people, we still need a fairly well-ordered and prosperous society. Unavoidably we're going to have to ban disorderly and destructive behaviour and enforce reasonable norms of honesty. Even the armies of Hell punish lying, theft, and unauthorised violence - or they would fall apart. To produce the society that maximises misery requires only a small but pervasive realignment of the one that produces most happiness.

If we try out other goals we find that whatever general entity we want to maximise, consequentialism will authorise much the same moral code. Certain negative qualities seem to be the only exceptions. What if we aim to maximise silence, for example? It seems unlikely that we want a bustling, prosperous society in that case: we might well want every living thing dead as soon as possible, and so embrace a very different code. But I think this result comes from the fact that negative goals like silence - the absence of noise - covertly change our principle from one of maximising to one of minimising, and that makes a real difference. Broadly, maximising anything yields the same moral code.

In fact, the vaguer we are about what we seek to maximise, the fewer the local distortions we are likely to get in the results. So it seems we should do best to go back now and replace the version of utilitarianism we took on provisionally with something we might call empty consequentialism, which simply enjoins us to choose actions that maximise our own legacy as agents, without tying us to happiness or any other specific desideratum. We should perform those actions which have the greatest consequences - that is, those that tend to produce the largest and most complex world.

We began by assuming that something was worth doing and have worked round to the view that everything is: or at least, that everything should be treated as worth doing. The moral challenge is simply to ensure our doing of things is as effective as possible. Looking at it that way reveals that even though we have moved away from the narrow specifics of the hypothetical imperatives we started with, we are still really in the same territory and still seeking to act effectively and coherently; it's just that we're trying to do so in a broader sense.

In fact what we're doing by seeking to maximise our consequential legacy is affirm and enlarge ourselves as persons. Personhood and agency are intimately connected. Acts, to deserve the name, must be intentional: things I do accidentally, unknowingly, or under direct constraint don't really count as actions of mine. Intentions don't exist in a free-floating state; they always

have their origin in a person; and we can indeed define a person as a source of intentions. We need not make any particular commitment about the nature of intentions, or about how they are generated. Whether the process is neural, computational, spiritual, or has some other nature, is not important here, so long as we can agree that in some significant sense new projects originate in minds, and that such minds are people. By adopting our empty consequentialism and the moral code it authorises, we are trying to imprint our personhood on the world as strongly as we can.

We live in a steadily unrolling matrix of cause and effect, each event following on from the last. If we live passively, never acting on intentions of our own, we never disturb the course of that process and really we do not have a significant existence apart from it. The more we act on projects of our own, the stronger and more vivid our personal existence becomes. The true weight of these original actions is measured by their consequences, and it follows that acting well in the sense developed above is the most effective way to enhance and develop our own personhood.

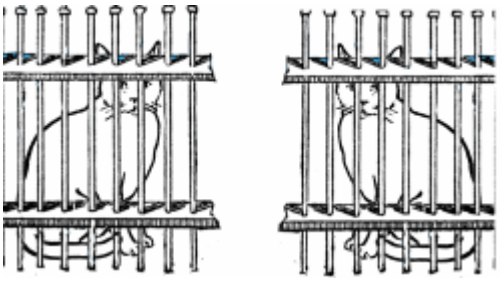
To me, this a congenial conclusion. Being able to root good behaviour and the observance of an ethical code in the realisation and enlargement of the self seems a satisfying result. Moreover, we can close the circle and see that this gives us at last some answer to the question we could not deal with at first - why be good? In the end there is no categorical imperative, but there is, as it were, a mighty hypothetical; if you want to exist as a person, and if you want your existence as a person to have any significance, you need to behave well. If you don't, then neither you nor anyone else need worry about the deeper reasons or ethics of your behaviour.

People who behave badly do not own the outcomes of their own lives; their behaviour results from the pressures and rewards that happen to be presented by the world. They themselves, as bad people, play little part in the shaping of the world, even when, as human beings, they participate in it. The first step in personal existence and personal growth is to claim responsibility and declare yourself, not merely reactive, but a moral being and an aspiring author of your own destiny. The existentialists, who have been sitting patiently smoking at a corner table, smile and raise an ironic eyebrow at our belated and imperfect enlightenment.

What about the people who rejected the constraints of morality and to greater or lesser degrees wanted to be left alone? Well, the system we've come up with enjoins us to adopt a moral code – but it leaves us to work out which one and explicitly allows for exceptions. Beyond that it consists of the general aspiration of 'empty consequentialism', but it is for us to decide how our consequential legacy is to be maximized. So the constraints are not tight ones. More important, it turns out that moral behaviour is the best way to escape from the tyranny of events and imprint ourselves on the world; obedience to the moral law, it turns out, is really the only way to be free.

Quantum Cognition

20 September 2015



We know all about the theory espoused by Roger Penrose and Stuart Hameroff, that consciousness is generated by a novel (and somewhat mysterious) form of quantum mechanics which occurs in the microtubules of neurons. For Penrose this helps explain how consciousness is non-computational, a result he has provided a separate proof for.

I have always taken it that this theory is intended to explain consciousness as we know it; but there are those who also think that quantum theory can provide a model which helps explain certain non-rational features of human cognition. This feature in the Atlantic nicely summarises what some of them say.

One example of human irrationality that might be quantumish is apparently provided by the good old Prisoner's Dilemma. Two prisoners who co-operated on a crime are offered a deal. If one rats, that one goes free while the other serves three years. If both rat, they serve two years; if neither do, they do one year. From a selfish point of view it's always better to rat, even though overall the no-rat strategy leads to least time in jail for the prisoners. Rationally, everyone should always rat, but in fact people quite often behave against their selfish interest by failing to do so. Why would that be?

Quantum theorists suggest that it makes more sense if we think of the prisoners being in a superposition of states between ratting and not-ratting, just as Schroedinger's cat superimposes life and death. Instead of contemplating the possible outcomes separately, we see them productively entangled (no, I'm not sure I quite get it, either).

There is of course another explanation; if the prisoners see the choice, not as a one-off, but as one in a series of similar trade-offs, the balance of advantage may shift, because those who are known to rat will be punished while those who don't may be rewarded by co-operation . Indeed, since people who seek to establish implicit agreements to co-operate over such problems will tend to do better overall in the long run, we might expect such behaviour to have positive survival value and be favoured by evolution.

A second example of quantum explanation is provided by the fact that question order can affect responses. There's an obvious explanation for this if one question affects the context by bringing something to the forefront of someone's mind. Asking someone whether they plan to drive home before asking them whether they want another drink may produce different results from asking the other way round for reasons that are not really at all mysterious. However, it's not always so clear cut and research demonstrates that a quantum model based on complementarity is really pretty good at making predictions.

How seriously are we to take this? Do we actually suppose that exotic quantum events in microtubules are directly responsible for the 'irrational' decisions? I don't know exactly how that would work and it seems rather unlikely. Do we go to the other extreme and assume that the quantum explanations are really just importing a useful model - that they are in fact ultimately metaphorical? That would be OK, except that metaphors typically explain the strange by invoking something understood. It's a little weird to suppose we could helpfully explain the

incomprehensible world of human motivation by appealing to the readily understood realm of quantum physics.

Perhaps it's best to simply see this as another way of thinking about cognition, something that surely can't be bad?

Hume The Buddhist

24 September 2015

Alison Gopnik suggests that David Hume was inspired by Buddhist ideas; she means well, but what nonsense! Hume is among the most important of Western philosophers and deserves to be widely appreciated, but if you wanted to strike a really damaging blow at popular understanding of what he says you could hardly do better than tangle him up with Buddha. Alas, the damage is probably done; the bogus linkage of the sceptical British empiricist with world-rejecting mysticism is probably lodged at the back of many minds.

Gopnik's piece here (the same ideas are set out in a paper here) suggests that Hume got an important idea - that the self is an illusion - from Buddhist thought. Her focus is narrow, with relatively little about the philosophy. Nearly all of her account is directed towards establishing the historical possibility of a particular route by which Hume could have come to hear of Buddhist doctrine, from Jesuits at La Flèche. She spends a lot of time on the details and parades the results as a success, but actually finds no evidence for anything more than the bare possibility that Hume might have, could have discussed Buddhism with someone who might have picked up a knowledge from someone else who had produced unpublished translations of certain texts and who had been at La Flèche some years earlier.

If Gopnik had found proof that Hume showed an interest in Buddhism or that it was ever mentioned to him at La Flèche her research might be of some value. As it is it's irrelevant, I think, because it's not all that unlikely that Hume could have found out about Buddhism anyway, from other sources, if he were at all interested. There is, of course, a vast historical chasm between could have and did. As one of the West's leading sceptics and the author of some of the most slyly biting sarcasm about religious beliefs it isn't particularly likely Hume would have sought to learn from Eastern religions, but let it pass; for the sake of argument we can grant that he might have heard the gist of Buddhist doctrine.

Was that the only place Hume could have got sceptical beliefs about the self? Well, no: there's a far more likely place. It was, after all, Descartes' most celebrated claim - then as now, one of the best-known theses of Western philosophy - that the existence of the self was the most certain thing we knew, and that it could be established simply by thinking about it. All we need do is negate that - and negating other people's claims is, after all, what philosophers do - and we're pretty much there. Descartes rests his whole system on his perception of the self; Hume comes along and says, when I think about it I find nothing there. Surely Hume, the commonsensical British empiricist, is inverting the celebrated foundation of the Frenchman's continental-style a priori reasoning?

Ah, but he could still have been influenced, couldn't he? Gopnik floats the idea that Hume could have forgotten about Buddhist ideas consciously while still having them working away in his subconscious mind. Such things do happen, and if the influence were subconscious it would

handily acquit Hume, the most honest and modest of men, from dishonesty or culpable silence over his sources.

The trouble is, Hume's source was explicitly his own mind, and it matters. He presents his view of the self, not as an interesting argument he heard somewhere that might be true, but as the direct result of his own inner observation. He was simply reporting how things looked to him. Was he wrong? What are we to say, that his introspections were systematically determined by his prior beliefs? That his unremembered conversation about Buddhism somehow rendered him incapable of seeing actual key features of his own mental landscape? Or that these same forgotten words enabled him to perceive an absence which his mind would otherwise have filled with a confabulated construct? Gopnik talks as though she supports Hume, but to discredit his introspection so radically would invalidate the grounds he is claiming; it would be a vigorous attack on Hume. It seems far simpler all round to believe that he was telling the simple truth: he looked into his mind and found no self there.

That is the real killer for me; Hume did not, in fact, say that the self was an illusion; he saw nothing. On this he may well be unique; he is surely original. Buddhists, and some modern philosophers, contend that the self is actually an illusion; a powerful one which it is hard to shake off, but one which is ultimately misleading. Hume, on the contrary, just saw no self.

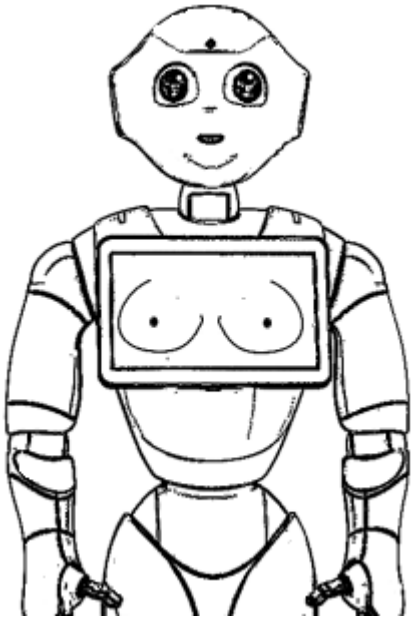
The distinction may not seem important, but it is; let me offer an analogy. Suppose we live near the great Nemonic Desert. Priests warn us that the fabled city of Nemonia, in the middle of the desert, is an hallucination. We will be tempted to stop and drink from its fountains, eat from the generous hospitality provided, and perhaps even stay; if we do we shall die, because the food and water are delusions and we'll die of thirst. Modern scientists have offered theories which explain how the mind constructs the delusion of the Nemonian city and why we should stop sending expeditions to look for it. In certain conditions our cognitive apparatus just constructs an encouraging but false perception for us, they say.

David Hume, on the other hand, tells us he walked across the desert keeping his eyes open and never saw anything but sand. Maybe it looks different to other people, he says, I can't argue with them about that; but to me there's just nothing there, simple as that.

To say then, that Hume got his disbelief in the city from listening to the priests is a dreadful error, a confusing misrepresentation, and really a bit of a slight against a man whose originality and honesty deserves better.

Sexbots

27 September 2015



We need to talk about sexbots. It seems (according to the Daily Mail - via MLU) that buyers of the new Pepper robot pal are being asked to promise they will not sex it up the way some naughty people have been doing; putting a picture of breasts on its touch screen and making poor Pepper tremble erotically when the screen is touched.

Just in time, some academics have launched the Campaign against Sex Robots. We've talked once or twice about the ethics of killbots; from thanatos we move inevitably to eros and the ethics of sexbots. Details of some of the thinking behind the campaign are set out in this paper by Kathleen Richardson of De Montfort University.

In principle there are several reasons we might think that sex with robots was morally dubious. We can put aside, for now at least, any consideration of whether it harms the robots emotionally or in any other way, though we might need to return to that eventually.

It might be that sex with robots harms the human participant directly. It could be argued that the whole business is simply demeaning and undignified, for example - though dignified sex is pretty difficult to pull off at the best of times. It might be that the human partner's emotional nature is coarsened and denied the chance to develop, or that their social life is impaired by their spending every evening with the machine. The key problem put forward seems to be that by engaging in an inherently human activity with a mere machine, the line is blurred and the human being imports into their human relationship behaviour only appropriate to robots: that in short, they are encouraged to treat human beings like machines. This hypothetical process resembles the way some young men these days are disparagingly described as "porn-educated" because their expectations of sex and a sexual relationship have been shaped and formed exclusively by what we used to call blue movies.

It might also be that the ease and apparent blamelessness of robot sex will act as a kind of gateway to worse behaviour. It's suggested that there will be "child" sexbots; apparently harmless in themselves but smoothing the path to paedophilia. This kind of argument parallels the ones about apparently harmless child porn that consists entirely of drawings or computer graphics, and so arguably harms no children.

On the other side, it can be argued that sexbots might provide a harmless, risk-free outlet for urges that would otherwise inconveniently be pressed on human beings. Perhaps the line won't really be blurred at all: people will readily continue to distinguish between robots and people, or perhaps the drift will all be the other way: no humans being treated as machines, but one or two machines being treated with a fondness and sentiment they don't really merit? A lot of people personalise their cars or their computers and it's hard to see that much harm comes of it.

Richardson draws a parallel with prostitution. That, she argues, is an asymmetrical relationship at odds with human equality, in which the prostitute is treated as an object: robot sex extends

and worsens that relationship in all respects. Surely it's bound to be a malign influence? There seem to be some problematic aspects to her case. A lot of human relationships are asymmetrical; so long as they are genuinely consensual most people don't seem bothered by that. It's not clear that prostitutes are always simply treated as objects: in fact they are notoriously required to fake the emotions of a normal sexual relationship, at least temporarily, in most cases (we could argue about whether that actually makes the relationship better or worse). Nor is prostitution simple or simply evil: it comes in many forms from many prostitutes who are atrociously trafficked, blackmailed and beaten, through those who regard it as basically another service job, through to some few idealistic practitioners who work in a genuine therapeutic environment. I'm far from being an advocate of the profession in any form, but there are some complexities and even if we accept the debatable analogy it doesn't provide us with a simple, one-size-fits-all answer.

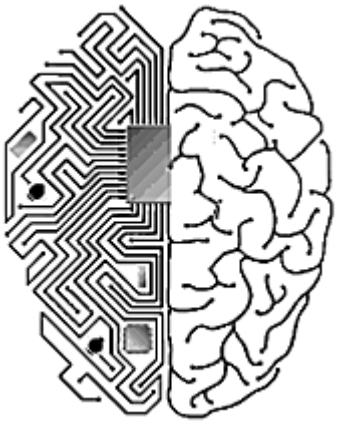
I do recognise the danger that the line between human and machine might possibly be blurred. It's a legitimate concern, but my instinct says that people will actually be fairly good at drawing the distinction and if anything robot sex will tend not to be thought of either as like sex with humans or sex with machines: it'll mainly be thought of as sex with robots, and in fact that's where a large part of the appeal will lie.

It's a bit odd in a way that the line-blurring argument should be brought forward particularly in a sexual context. You'd think that if confusion were to arise it would be far more likely and much more dangerous in the case of chat-bots or other machines whose typical interactions were relatively intellectual. No-one, I think, has asked for Siri to be banned.

My soggy conclusion is that things are far more complex than the campaign takes them to be, and a blanket ban is not really an appropriate response.

Equivocating Uploads

4 October 2015



Information about the brain is not the same as information in the brain; yet in discussions of mind uploading, brain simulation, and mind reading the two are quite often conflated or confused. Equivocating between the two makes the task seem far easier than it really is. Scanners of various kinds exist, after all, and have greatly improved in recent years; technology usually goes on improving over time. If all we need is to get a really good scan of the brain in order to understand it, then surely it can only be a matter of time? Alas, information about the brain is an inadequate proxy for the information in the brain that we really need.

We're often asked to imagine a scanner that can read off the structural details of a brain down to any required level of detail. Usually we're to assume this can be done non-destructively, or even without disturbing the content and operation of the brain at all. These are of course unlikely feats, not just beyond existing technology but rather hard to imagine even on the most optimistic view of future developments. Sometimes the ready confidence that this miracle will one day be within our grasp is so poorly justified I find it tempting to think that the belief in the possibility of such magic scans is being buoyed up not by sanguine technological speculation but unconsciously by much older patterns of thinking; that the mind is located in breath, or airy spirits, or some other immaterial substance that can be sucked out of a body and replaced without physical consequences. Of course on the other side it's perfectly true that lots of things once considered impossible are now routinely achieved.

But even allowing ourselves the most miraculous knowledge of the brain's structure, so what? We could have an exquisite plan of the structure of a disk or a book without knowing what story it contained. Indeed, it would only take a small degree of inaccuracy or neglect in copying to leave us with a duplicate that strongly resembled the original but actually reproduced none of the information bearing elements; a disk with nothing on it, a book with random ink patterns.

Yeah but, the optimists say; the challenge may be huge, the level of detail required orders of magnitude beyond anything previously attempted, but if we copy something with enough fidelity the information just is going to come along with the copy necessarily. A perfect copy just has to include a perfect copy of the information. Granted, in the case of a book it's not much use if we have the information but don't know how to read it. The great thing about simulating a brain, though, is that we don't even need to understand. We can just set it up and start it going. We may never know directly what the information in the brain was, but it'll do its job; the mind will upload, the simulation will run.

In the case of mind reading the marvellous flexibility of the mind also offers us a chance to cheat by taking some measurable, controllable brain function and simply using it as a signalling device. It works, up to a point, but it isn't clear why brain communication by such lashed-up indirect routes is any more telepathy than simply talking to someone; in both cases the actual information in the brain remains inaccessible except through a deliberate signalling procedure.

Now of course a book or a disk is in some important ways actually a far simpler challenge than a brain. The people who designed, made, and use the disk or the book take great care to ensure

that a specific, readily readable set of properties encodes the information required in a regular, readable form. These are artefacts designed to carry information, as is a computer. The brain is not artefactual and does not need to be legible. There's no need for a clear demarcation between information-bearing elements and the rest, and there's no need for a standardised encoding or intelligible structures. There are, in fact many complex elements that might have a role in holding information.

Suppose we recalibrated our strategy and set out to scan just the information in the brain; what would we target? The first candidate these days is the connectome; the overall pattern of neural connections within the brain. There's no doubt this kind of research is currently very lively and interesting - see for example this recent study. Current research remains pretty broad brush stuff and it's not really clear how much detail will ever be manageable; but what if we could map the connections perfectly? How could we read off the content? It's actually highly unlikely that all the information in the brain is encoded as properties of a network. The functional state of a neuron depends on many things, in particular the receptors and transmitters; the large known repertoire of these has greatly increased in recent years. We know that the brain does not operate simply through electrical transmission, with chemical controls from the endocrine system and elsewhere playing a large and subtle part. It's not at all unlikely that astrocytes, the non-neural cells in the brain, have a significant role in modulating and even controlling its activity. It's not at all unlikely, on the other hand, that ephaptic coupling or other small electrical induction effects have a significant role, too. And while myself I wouldn't place any bets on exotic quantum physics being relevant, as some certainly believe, I think it would be very rash to assume that biology has no further tricks up its sleeve in the shape of important mechanisms we haven't even noticed yet.

None of that can currently be ruled out of court as irrelevant. A computer has a specified way of working and if electrical interference starts changing the value of some bits in working memory you know it's a bug, not a feature. In the brain, it could be either; the only way to judge is whether we like the results or not. There's no reason why astrocyte states, say, can't be key for one brain region or for one personality, and irrelevant for others, or even legitimate at one time and unhelpful interference at others. We just cannot know what to point our magic scanner at, and it may well be that the whole idea of information recorded in but distinct from a substrate just isn't appropriate.

Yeah but again, total perfect copy? In principle if we get everything, we get everything, don't we? The problem is that we can't have everything. Copying, simulating, or transmitting necessarily involve transitions during which some features are unavoidably left out. Faith in the possibility of a good copy rests on the belief that we can identify a sufficient set of relevant features; so long as those are preserved, we're good. We're optimistic that one day we can do a job on the physical properties which is good enough. But that's just information about the brain.

Markram's Electric Gland

12 October 2015

The brain is not a gland. But Henry Markram seems to be in danger of simulating one - a kind of electric gland.

What am I on about? The Blue Brain Project has published details of its most ambitious simulation yet; a computer model of a tiny sliver of rat brain. That may not sound exciting on the face of it, but the level of detail is unprecedented...

> *The reconstruction uses cellular and synaptic organizing principles to algorithmically reconstruct detailed anatomy and physiology from sparse experimental data. An objective anatomical method defines a neocortical volume of 0.29 ± 0.01 mm3 containing ~31,000 neurons, and patch-clamp studies identify 55 layer-specific morphological and 207 morpho-electrical neuron subtypes. When digitally reconstructed neurons are positioned in the volume and synapse formation is restricted to biological bouton densities and numbers of synapses per connection, their overlapping arbors form ~8 million connections with ~37 million synapses.*

The results are good. Without parameter tuning – that is, without artificial adjustments to make it work the way it should work – the digital simulation produces patterns of electrical activity that resemble those of real slivers of rat brain. The paper is accessible here. It seems a significant achievement and certainly attracted a lot of generally positive attention - but there are some significant problems. The first is that the methodological issues which were always evident remain unresolved. The second is certain major weaknesses in the simulation itself. The third is that as a result of these weaknesses the simulation implicitly commits Markram to some odd claims, ones he probably doesn't want to make.

First, the methodology. The simulation is claimed as a success, but how do we know? If we're simulating a heart, then it's fairly clear it needs to pump blood. If we're simulating a gland, it needs to secrete certain substances. The brain? It's a little more complicated. Markram seems implicitly to take the view that the role of brain tissue is to generate certain kinds of electrical activity; not particular functional outputs, just generic activity of certain kinds.

One danger with that is a kind of circularity. Markram decides the brain works a certain way, he builds a simulation that works like that, and triumphantly shows us that his simulation does indeed work that way. Vindication! It could be that he is simply ignoring the most important things about neural tissue, things that he ought to be discovering. Instead he might just be cycling round in confirmation of his own initial expectations. One of the big dangers of the Blue Brain project is that it might entrench existing prejudices about how the brain works and stop a whole generation from thinking more imaginatively about new ideas.

The Blue simulation produces certain patterns of electrical activity that look like those of real rat brain tissue – but only in general terms. Are the actual details important? After all, a string of random characters with certain formal constraints looks just like meaningful text, or valid code,

or useful stretches of DNA, but is in fact useless gibberish. Putting in constraints which structure the random text a little and provide a degree of realism is a relatively minor task; getting output that's meaningful is the challenge. It looks awfully likely that the Blue simulation has done the former rather than the latter, and to be brutal that's not very interesting. At worst it could be like simulating an automobile whose engine noise is beautifully realistic but never moves. We might well think that the project is falling into the trap I mentioned last time: mistaking information about the brain for the information in the brain.

Now it could be that actually the simulation is working better than that; perhaps it isn't as generic as it seems, perhaps this particular bit of rat brain works somewhat generically anyway; or perhaps somehow in situ the tissue trains or conditions itself, saving the project most of the really difficult work. The final answer to the objections above might be if the simulation could be plugged back into a living rat brain and the rat behaviour shown to continue properly. If we could do that it would sidestep the difficult questions about how the brain operates; if the rat behaves normally, then even though we still don't know why, we know we're doing it right. In practice it doesn't seem very likely that that would work, however: the brain is surely about producing specific control outputs, not about glandularly secreting generic electrical activity.

A second set of issues relates to the limitations of the simulation. Several of the significant factors I mentioned above have been left out; notably there are no astrocytes and no neurotransmitters. The latter is a particularly surprising omission because Markram himself has in the past done significant work on trying to clarify this area in the past. The fact that the project has chosen to showcase a simulation without them must give rise to a suspicion that its ambitions are being curtailed by the daunting reality; that there might even be a dawning realisation internally that what has been bitten off here really is far beyond chewing and the need to deliver has to trump realism. That would be a worrying straw in the wind so far as the project's future is concerned.

In addition, while the simulation reproduces a large number of different types of neuron, the actual wiring has been determined by an algorithm. A lot depends on this: if the algorithm generates useful and meaningful connections then it is itself a master-work beside which the actual simulation is trivial. If not, then we're back with the question of whether generic kinds of connection are really good enough. They may produce convincing generic activity, and maybe that's even good enough for certain areas of rat brain, but we can be pretty sure it isn't good enough for brain activity generally.

Harking back for a moment to methodology, there's still something odd in any case about trying to simulate a process you don't understand. Any simulation reproduces certain features of the original and leaves others out. The choice is normally determined by a thesis about how and why the thing works: that thesis allows you to say which features are functionally necessary and which are irrelevant. Your simulation only models the essential features and its success therefore confirms your view about what really matters and how it all operates. Markram, though, is not starting with an explicit thesis. One consequence is that it is hard to tell whether his simulation is a success or not, because he didn't tell us clearly enough in advance what it was he was trying to make happen. What we can do is read off the implicit assumptions that the project cannot help embodying in its simulation. To hold up the simulation as a success is to make the implicit claim that the brain is basically an electrical network machine whose essential function is to generate certain general types of neural activity. It implicitly affirms that the features left out of the simulation – notably the vast array and complex role of neural transmitters and receptors – are essentially irrelevant. That is a remarkable claim, quite unlikely,

and I don't think it's one Markram really wants to make. But if he doesn't, consistency obliges him to downplay the current simulation rather than foreground it.

To be fair, the simulation is not exactly being held up as a success in those terms. Markram describes it as a first draft. That's fair enough up to a point (except that you don't publish first drafts), but if our first step towards a machine that wrote novels was one that generated the Library of Babel (every possible combination of alphabetic characters plus spaces) we might doubt whether we were going in the right direction. The Blue Brain project in some ways embodies technological impatience; let's get on and build it and worry about the theory later. The charge is that as a result the project is spending its time simulating the wrong features and distracting attention from the more difficult task of getting a real theoretical understanding; that it is making an electric gland instead of a brain.

Four Kinds Of Hard

18 October 2015

Not one Hard Problem, but four. Jonathan Dorsey, in the latest JCS, says that the problem is conceived in several different ways and we really ought to sort out which we're talking about.

The four conceptions, rewritten a bit by me for what I hope is clarity, are that the problem is to explain why phenomenal consciousness:

1. ...arises from the physical (using only a physicalist ontology)
2. ...arises from the physical, using any ontology
3. ...arises at all (presumably from the non-physical)
4. ...arises at all or cannot be explained.

I don't really see these as different conceptions of the problem (which simply seems to be the explanation of phenomenal consciousness), but rather as different conceptions of what the expected answer is to be. That may be nit-picking; useful distinctions in any case. Dorsey offers some pros and cons for each of the four.

In favour of number one, it's the most tightly focused. It also sits well in context, because Dorsey sees the problem as emerging under the dominance of physics. The third advantage is that it confines the problem to physicalism and so makes life easy for non-physicalists (not sure why this is held to be one of the pros, exactly). Against; well, maybe that context is dominating too much? Also the physicalist line fails to acknowledge Chalmers' own naturalist but non-physicalist solution (it fails to acknowledge lots of other potential solutions too, so I'm not quite clear why Chalmers gets this special status at this point - though of course he did play a key role in defining the Hard Problem).

Number two's pros and cons are mostly weaker versions of number one's. It too is relatively well-focused. It does not identify the Hard Problem with the Explanatory Gap (that could be a con rather than a pro in my humble opinion). It fits fairly well in context and makes life relatively easy for traditional non-physicalists. It may yield a bit too much to the context of physics and it may be too narrow.

Number three has the advantage of focusing on the basics, and Dorsey thinks it gives a nice clear line between Hard and Easy problems. It provides a unifying approach - but it neglects the physical, which has always been central to discussion.

Number four provides a fully extended version of the problem, and makes sense of the literature by bringing in eliminativism. In a similar way it gives no-one a free pass; everyone has to address it. However, in doing so it may go beyond the bounds of a single problem and extend the issues to a wide swathe of philosophy of mind.

Dorsey thinks the answer is somewhere between 2 and 3; I'm more inclined to think it's most likely between 1 and 2.

Let's put aside the view that phenomenal consciousness cannot be explained. There are good arguments for that conclusion, but to me they amount to opting out of a game which is by no means clearly lost. So the problem is to explain how phenomenal consciousness arises. The explanation surely has to fit into some ontology, because we need to know what kind of thing phenomenal experience really is. My view is that the high-level differences between ontologies actually matter less than people have traditionally thought. Look at it this way: if we need an ontology, then it had better be comprehensive and consistent. Given those two properties, we might as well call it a monism. because it encompasses everything and provides one view, even if that one view is complex.

So we have a monism: but it might be materialism, idealism, spiritualism, neutral monism, or many others. Does it matter? The details do matter, but if we've got one substance it seems to me it doesn't matter what label we apply. Given that the material world and its ontology is the one we have by far the best knowledge of, we might as well call it materialism. It might turn out that materialism is not what we think, and it might explain all sorts of things we didn't expect it to deal with, but I can't see any compelling reason to call our single monist ontology anything else.

So what are the details, and what ontology have we really got? I'm aware that most regulars here are pretty radical materialists, with some exceptions (hat tip to cognicious); people who have some difficulty with the idea that the cosmos has any contents besides physical objects; uncomfortable with the idea of ideas (unless they are merely conjunctions of physical objects) and even with the belief that we can think about anything that isn't robustly physical (so much for mathematics...). That's not my view. I'm also a long way from from being a Platonist, but I do think the world includes non-physical entities, and that that doesn't contradict a reasonable materialism. The world just is complex and in certain respects irreducible; probably because it's real. Reduction, maybe, is essentially a technique that applies to ideas and theories: if we can come up with a simpler version that does the job, then the simpler version is to be adopted. But it's a mistake to think that that kind of reduction applies to reality itself; the universe is not obliged to conform to a flat ontology, and it does not. At the end of the day - and I say this with the greatest regret - the apprehension of reality is not purely a matter of finding the simplest possible description.

I believe the somewhat roomier kind of materialism I tend to espouse corresponds generally with what we should recognise as the common sense view, and this yields what might be another conception of the Hard Problem...

5. ...arises from the physical (in a way consistent with common sense).

Eliminating Common Sense

25 October 2015

The more you know about the science of the mind, the less appealing our common sense ideas seem. Ideas about belief and desire motivating action just don't seem to match up in any way with what you see going on. So, at least, says Jose Luis Bermudez in Arguing for Eliminativism (freely available on Academia, but you might need to sign in). Bermudez sympathises with Paul Churchland's wish to sweep the whole business of common sense psychology away; but he wants to reshape Churchland's attack, standing down the 'official' arguments and bringing forward others taken from within Churchland's own writing on the subject.

Bermudez sketches the complex landscape with admirable clarity. He notes Boghossian has argued that eliminativism of this kind is incoherent: but Boghossian construed eliminativism as an attack on all forms of content. Bermudez has no desire to be so radical and champions a purely psychological eliminativism.

If something's wrong with common sense psychology it could either be that what it says is false, or that what it says is not even capable of being judged true or false. In the latter case it could, for example, be that all common sense talk of mental states is nothing more than a complex system of reflexive self-expression like grunts and moans. Bermudez doesn't think it's like that: the propositions of common sense psychology are meaningful, they just happen to be erroneous.

It therefore falls to the eliminativist to show what the errors are. Bermudez has a two-horned strategy: first, we can argue that as a matter of fact, we don't rely on common sense understanding as much as we think. Second, we can look for ways to show that the kind of propositional content implied by common sense views is just incompatible with the mechanism that actually underlie human action and behaviour as revealed by scientific investigation.

There are, in fact, two different ways of construing common sense psychology. One is that our common sense understanding is itself a kind of theory of mind: this is the 'theory theory' line. To disprove this we might try to bring out what the common sense theory is and then attack it. The other way of construing common sense is that we just use our own minds as a model: we put ourselves in the other person's shoes and imagine how we should think and react. To combat this one we should need a slightly different approach; but it seems Bermudez's strategy is good either way.

I think the first horn of the attack works better than the second - but not perfectly. Bermudez rightly says it is very generally accepted that to negotiate complex social interactions we need to ascribe beliefs and desires to other people and draw conclusions about their likely behaviour. It ain't necessarily so. Bermudez quotes the Prisoner's Dilemma, the much-cited example where we have been arrested: if we betray our partner in crime we'll get better terms whatever the

other one does. We don't, Bermudez points out, need to have any particular beliefs about what the other person will do: we can work out the strategy from just knowing the circumstances.

More widely, Bermudez contends, we often don't really need to know what an individual has in mind. If we know that person is a butcher, or a waiter, then the relevant social interaction can be managed without any hypothesising about beliefs and desires. (In fact we can imagine a robot butcher/waiter who would certainly lack any beliefs or desires but could execute the transactions perfectly well.)

That is fine as far as it goes, but it isn't that hard to think of examples where the ascription of beliefs seems relevant. In particular, the interpretation of speech, especially the reading of Gricean implicatures, seems to rely on it. Sometimes it also seems that the ascription of emotional states is highly relevant, or hypotheses about what another person knows, something Bermudez doesn't address.

It's interesting to reflect on what a contrast this is with Dennett. I think of Dennett and Churchland as loosely allied: both sceptics about qualia, both friendly to materialist, reductive thinking. Yet here Bermudez presents a Churchlandish view which holds that ascriptions of purpose are largely useless in dealing with human interaction, while Dennett's Intentional Stance of course requires that they are extremely useful.

Bermudez doesn't think this kind of argument is sufficient, anyway, hence the second horn in which he tries to sketch a case for saying that common sense and neurons don't fit well. The real problem here for Bermudez is that we don't really know how neurons represent things. He makes a case for kinds of representation other than the sort of propositional representation he thinks is required by the standard common sense view (ie, we believe or desire that xxx...). It's true that a mess of neurons doesn't look much like a set of well-formed formulae, but to cut to the chase I think Bermudez is pursuing a vain quest. We know that neurons can deal with ascriptions of propositional belief and desire (otherwise how would we even be able to think and talk about them) so it's not going to be possible to rule them out neurologically. Bermudez presents some avenues that could be followed, but even he doesn't seem to think the case can be clinched as matters stand.

I wonder if he needs to? It seems to me that the case for elimination does not rest on proving the common sense concepts false, only on their being redundant. If Bermudez can show that all ascriptions of belief and desire can for practical purposes be cashed out or replaced by cognition about the circumstances and game-theoretic considerations, then simple parsimony will get him the elimination he seeks.

He would still, of course, be left with explaining why the human mind adopts a false theory about itself instead of the true one: but we know some ways of explaining that - for example, ahem, through the Blindness of the Brain (ie that we're trapped within our limitations and work with the poor but adequate heuristics gifted to us, or perhaps foisted on us, by evolution).

Baby Solipsists

1 November 2015

Are babies solipsists? Ali, Spence and Bremner at Goldsmiths say their recent research suggests that they are “tactile solipsists”.

To be honest that seems a little bit of a stretch from the actual research. In essence this tested how good babies were at identifying the location of a tactile stimulus. The researchers spent their time tickling babies and seeing whether the babies looked in the direction of the tickle or not (the life of science is tough, but somebody’s got to do it). Perhaps surprisingly, perhaps not, the babies were in general pretty good at this. In fact the youngest ones were less likely to be confused by crossing their legs before tickling their feet, something that reduced the older ones’ success rate to chance levels, and in fact impairs the performance of adults too.

The reason for this is taken to be that long experience leads us to assume a stimulus to our right hand will match an event in the right visual field, and so on. After the correlations are well established the brain basically stops bothering to check and is then liable to be confused when the right hand (or foot) is actually on the left, or vice versa.

This reminded me a bit of something I noticed with my own daughters: when they were very small their fingers all worked independently and were often splayed out, with single fingers moving quite independently; but in due course they seemed to learn that not much is achieved in most circumstances by using the four digits separately and that you might as well use them in concert by default to help with grasping, as most of us mostly do except when using a keyboard.

Very young babies haven’t had time to learn any of this and so are not confused by laterally inconsistent messages. The Goldsmiths’ team read this as meaning that they are in essence just aware of their own bodies, not aware of them in relation to the world. It could be so, but I’m not sure it’s the only interpretation. Perhaps it’s just not that complex.

There are other reasons to think that babies are sort of solipsistic. There’s some suggestive evidence these days that babies are conscious of their surroundings earlier than we once thought, but until recently it’s been thought that self-awareness didn’t dawn until around fifteen months, with younger babies unaware of any separation between themselves and the world. This was partly based on the popular mirror test, where a mark is covertly put on the subject’s face. When shown themselves in a mirror, some touch the mark; this is taken to show awareness that the reflection is them, and hence a clear sign of self awareness. The test has been used to indicate that such self-awareness is mainly a human thing, though also present in some apes, elephants, and so on.

The interpretation of the mirror test always seemed dubious to me. Failure to touch your own face might not mean you’ve failed to recognise yourself; contrariwise, you might think the

reflection was someone else but still be motivated to check your own face to see whether you too had a mark. If people out there are getting marked, wouldn't you want to check?

Sure enough about five years ago evidence emerged that the mirror test is in fact very much affected by cultural factors and that many human beings outside the western world react quite differently to a mirror. It's not all that surprising that if you've seen people use mirrors to put on make-up (or shave) regularly your reactions to one might be affected. If we were to rely on the mirror test, it seems many Kenyan six-year-olds would be deemed unaware of their own existence.

Of course the question is in one sense absurd: to be any kind of solipsist is, strictly speaking, to hold an explicit philosophical position which requires quite advanced linguistic and conceptual apparatus which small infants certainly don't have. For the question to be meaningful we have to have a clear view about what kinds of beliefs babies can be said to hold. I don't doubt that hold some inexplicit ones, and that we go on holding beliefs in the same way alongside others at many different levels. If we reach out to catch a ball we can in some sense be said to hold the belief that it is following a certain path, although we may not have entertained any conscious thoughts on the matter. At the other end of the spectrum, where we solemnly swear to tell the truth, the whole truth, and nothing but the truth, the belief has been formulated with careful specificity and we have (one hopes) deliberated inwardly at the most abstract levels of thought about the meaning of the oath. The complex and many-layered ways in which we can believe things have yet to be adequately clarified I think; a huge project and since introspection is apparently the only way to tackle it, a daunting one.

For me the only certain moral to be drawn from all the baby-tickling is one which philosophers might recognise: the process of learning about the world is at root a matter of entering into worse and grander confusions.

Are We Aware Of Concepts?

8 November 2015

Are ideas conscious at all? Neuroscience of Consciousness is a promising new journal from OUP introduced by the editor Anil Seth here. It has an interesting opinion piece from David Kemmerer which asks - are we ever aware of concepts, or is conscious experience restricted to sensory, motor and affective states?

On the face of it a rather strange question? According to Kemmerer there are basically two positions. The 'liberal' one says yes, we can be aware of concepts in pretty much the same kind of way we're aware of anything. Just as there is a subjective experience when we see a red rose, there is another kind of subjective experience when we simply think of the concept of roses. There are qualia that relate to concepts just as there are qualia that relate to colours or smells, and there is something it is like to think of an idea. Kemmerer identifies an august history for this kind of thinking stretching back to Descartes.

The conservative position denies that concepts enter our awareness. While our behaviour may be influenced by concepts, they actually operate below the level of conscious experience. While we may have the strong impression that we are aware of concepts, this is really a mistake based on awareness of the relevant words, symbols, or images. The intellectual tradition behind this line of thought is apparently a little less stellar - Kemmerer can only push it back as far as Wundt - but it is the view he leans towards himself.

So far so good - an interesting philosophical/psychological issue. What's special here is that in line with the new journal's orientation Kemmerer is concerned with the neurological implications of the debate and looks for empirical evidence. This is an unexpected but surely commendable project.

To do it he addresses three particular theories. Representing the liberal side he looks at Global Neural Workspace Theory (GNWT) as set out by Dehaene, and Tononi's Integrated information Theory (IIT)' on the conservative side he picks the Attended Intermediate-Level Representation Theory (AIRT) of Prinz. He finds that none of the three is fully in harmony with the neurological evidence, but contends that the conservative view has distinct advantages.

Dehaene points to research that identified specific neurons in a subject's anterior temporal lobes that fire when the subject is shown a picture of, say, Jennifer Aniston (mentioned on CE - rather vaguely). The same neuron fires when shown photographs, drawing, or other images, and even when the subject is reporting seeing a picture of Aniston. Surely then, the neuron in some sense represents not an image but the concept of Jennifer Aniston? against theconservative view Kemmerer argues that while a concept may be at work, imagery is always present in the conscious mind; indeed, he contends, you cannot think of 'Anistonicity' in itself without a particular image of Aniston coming to mind. Secondly he quotes further research which shows that deterioration of this portion of the brain impairs our ability to recognise, but not to see,

faces. This, he contends, is good evidence that while these neurons are indeed dealing with general concepts at some level, they are contributing nothing to conscious awareness, reinforcing the idea that concepts operate outside awareness. According to Tononi we can be conscious of the idea of a triangle, but how can we think of a triangle without supposing it to be equilateral, isosceles, or scalene?

Turning to the conservative view, Kemmerer notes that AIRT has awareness at a middle level, between the jumble of impressions delivered by raw sensory input on the one hand, and the invariant concepts which appear at the high level. Conscious information must be accessible but need not always be accessed. It is implemented as gamma vector waves. This is apparently easier to square with the empirical data than the global workspace, which implies that conscious attention would involve a shift into the processing system in the lateral prefrontal cortex where there is access to working memory - something not actually observed in practice. Unfortunately although the AIRT has a good deal of data on its side the observed gamma responses don't in fact line up with reported experience in the way you would expect if it's correct.

I think the discussion is slightly hampered by the way Kemmerer uses 'awareness' and 'consciousness' as synonyms. I'd be tempted to reserve awareness for what he is talking about, and allow that concepts could enter consciousness without our being (subjectively) aware of them. I do think there's a third possibility being overlooked in his discussion - that concepts are indeed in our easy-problem consciousness while lacking the hard-problem qualia that go with phenomenal experience. Kemmerer alludes to this possibility at one point when he raises Ned Block's distinction between access and phenomenal (a- and p-consciousness), but doesn't make much of it.

Whatever you think of Kemmerer's ambivalent;y conservative conclusion, I think the way the paper seeks to create a bridge between the philosophical and the neurological is really welcome and, to a degree, surprisingly successful. If the new journal is going to give us more like that it will definitely be a publication to look forward to.

Conservativising Consciousness

16 November 2015

This paper from Chawke and Kanai reports unexpected effects on subjects' political views, brought about by stimulation of the dorsolateral prefrontal cortex (DLPFC). It seems to make people more conservative.

The research set out to build on earlier studies. Those seemed to suggest that the DLPFC had a role in flagging up conflicts; noting for us where the evidence was suggesting our views might need to be changed. Generally people stick to a particular outlook (the researchers suggest that avoidance of cognitive dissonance and similar stresses is one reason) but every now and then a piece of evidence comes along that suggest we really have to do a little bit of reshaping, and the DLPFC helps with that unwelcome process.

If that theory is right. then gingering up the DLPFC ought to make people readier to change their existing views. To test this, the authors set up arrangements to deliver random trans-cranial noise stimulation bilaterally to the relevant areas. They tested subjects' political views beforehand; showed them a party political broadcast, and then checked to see whether the subjects' views had in fact changed.

This was at Sussex, so the political framework was a British one of Labour versus Conservative. The expectation was that stimulating the DLPFC would make the subjects more receptive to persuasion and so more inclined to adjust their views slightly in response to what they were seeing; so Labour-inclined subjects would move to the right while conservative-inclined ones moved to the left.

Briefly, that isn't what happened: instead there was a small but significant general shift to the right. Why could that be? To be honest it's impossible to say, but hypothetically we might suppose that the DLPFC is not, after all, responsible for helping us change our view in the fare of contrary evidence, but simply a sceptical or disbelieving module that allows us to doubt or discard political opinions. Arguably - and I hope I'm not venturing into controversial territory - right wing views tend to correspond with general doubt about political projects and a feeling that things are best left alone; we could say that the fewer politics you have the more you tend to be on the right?

Whether that's true or not it seems alarming that stimulating the brain directly with random noise can affect your political views; it suggests an unscrupulous government could change the result of an election by irradiating the poling stations.

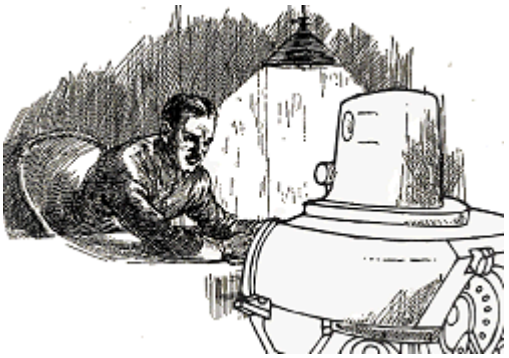
What did it feel like for the subjects? Nothing, it seems; the experimenters were careful to ensure that control subjects got the same kind of experience although their DLPFC was left alone. Subjects were apparently unaware of any change in their views (and we're only talking shifts on a small scale, not Damascene conversions to the opposite party).

Perhaps in the end it's not quite as alarming as it seems. Suppose we played our subjects bursts of ordinary random acoustic noise? That would be rather irritating; it might make them overall a

little angrier and less patient - might that not also have a small temporary effect on their voting pattern...?

Conversation With A Zombie

23 November 2015



Tom Clark has written a nice imaginary dialogue with a zombie on the subject of qualia. Could we in fact learn useful lessons from talking to a robot which lacked qualia?

Perhaps not; one view would be that since the robot's mind presumably works in the same way as ours, it would have similar qualia: or would think it did. We know that David Chalmers' zombie twin talked and philosophised about its qualia in exactly the same way as the original.

It depends on what you mean by qualia, of course. Some people conceive of qualia as psychological items that add extra significance or force to experience; or as flags that draw attention to something of potential interest. Those play a distinct role in decision making and have an influence on behaviour. If robots were really to behave like us, they would have to have some functional analogue of that kind of qualia, and so we might indeed find that talking to them on the subject was really no better or worse than talking to our fellow human beings.

But those are not real qualia, because they are fully naturalised and effable things, measurable parts of the physical world. Whether you are experiencing the same blue quale as me would, if these flags or intensifiers were qualia, be an entirely measurable and objective question, capable of a clear answer. Real, philosophically interesting qualia are far more slippery than that.

So we might expect that a robot would reproduce the functional, a-consciousness parts of our mind, and leave the phenomenal, p-consciousness ones out. Like Tom's robot they would presumably be puzzled by references to subjective experience. Perhaps, then, there might be no point in talking to them about it because they would be constitutionally incapable of shedding any light on it. they could tell us what the zombie life is like, but don't we sort of know that already? They could play the kind of part in a dialogue that Socrates' easily-bamboozled interlocutors always seemed to do, but that's about it, presumably?

Or perhaps they would be able to show us, by providing a contrasting example, how and why it is that we come to have these qualia? There's something distinctly odd about the way qualia are apparently untethered from physical cause and effect, yet only appear in human beings with their complex brains. Or could it be that they're everywhere and it's not that only we have them, it's more that we're the only entities that talk about them (or about anything)?

Perhaps talking to a robot would convince us in the end that in fact, we don't have qualia either: that they are just a confused delusion. One scarier possibility though, is that robots would understand them all too well.

"Oh," they might say, "Yes, of course we have those. But scanning through the literature it seems to us you humans only have a very limited appreciation of the qualic field. You experience simple local point qualia, but you have no perception of higher-order qualia; the qualia of the surface or the solid, or the complex manifold that seems so evident to us. Gosh, it must be *awful*!

Robots On Drugs

29 November 2015

Over the years many variants and improvements to the Turing Test have been proposed, but surely none more unexpected than the one put forward by Andrew Smart anticipating his forthcoming book *Beyond Zero and One*. He proposes that in order to be considered truly conscious, a robot must be able to take an acid trip.

He starts out by noting that computers seem to be increasing in intelligence (whatever that means), and that many people see them attaining human levels of performance by 2100 (actually quite a late date compared to the optimism of recent decades; Turing talked about 2000, after all). Some people, indeed, think we need to be concerned about whether the powerful AIs of the future will like us or behave well towards us. In my view these worries tend to blur together two different things; improving processing speeds and sophistication of programming on the one hand, and transformation from a passive data machine into a spontaneous agent, quite a different matter. Be that as it may, Smart reasonably suggests we could give some thought to whether and how we should make machines conscious.

It seems to me - this may be clearer in the book - that Smart divides things up in a slightly unusual way. I've got used to the idea that the big division is between access and phenomenal consciousness, which I take to be the same distinction as the one defined by the terminology of Hard versus Easy Problems. In essence, we have the kind of consciousness that's relevant to behaviour, and the kind that's relevant to subjective experience.

Although Smart alludes to the Chalmersian zombies that demonstrate this distinction, I think he puts the line a bit lower; between the kind of AI that no-one really supposes is thinking in a human sense and the kind that has the reflective capacities that make up the Easy Problem. He seems to think that experience just goes with that (which is a perfectly viable point of view). He speaks of consciousness as being essential to creative thought, for example, which to me suggests we're not talking about pure subjectivity.

Anyway, what about the drugs? Smart sedans to think that requiring robots to be capable of an acid trip is raising the bar, because it is in these psychedelic regions that the highest, most distinctive kind of consciousness is realised. He quotes Hofman as believing that LSD...

> *...allows us to become aware of the ontologically objective existence of consciousness and ourselves as part of the universe...*

I think we need to be wary here of the distinction between becoming aware of the universal ontology and having the deluded feeling of awareness. We should always remember the words of Oliver Wendell Holmes Sr:

> *...I once inhaled a pretty full dose of ether, with the determination to put on record, at the earliest moment of regaining consciousness, the thought I should find uppermost in my mind. The mighty music of the triumphal march into nothingness reverberated through my brain, and filled me with a sense of infinite possibilities, which made me an archangel for the moment. The veil of eternity was lifted. The one great truth which*

> *underlies all human experience, and is the key to all the mysteries that philosophy has sought in vain to solve, flashed upon me in a sudden revelation. Henceforth all was clear: a few words had lifted my intelligence to the level of the knowledge of the cherubim. As my natural condition returned, I remembered my resolution; and, staggering to my desk, I wrote, in ill-shaped, straggling characters, the all-embracing truth still glimmering in my consciousness. The words were these (children may smile; the wise will ponder): "A strong smell of turpentine prevails throughout…*

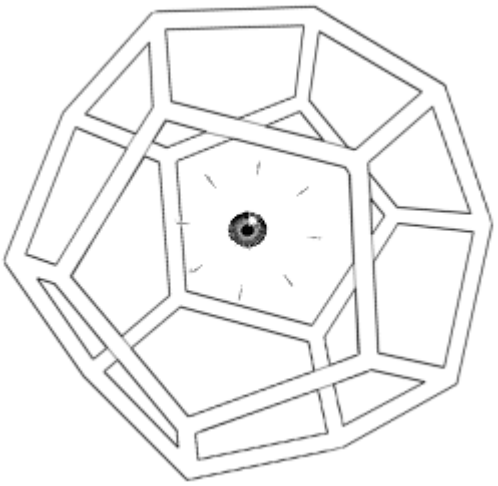
A second problem is that Smart believes (with a few caveats) that any digital realisation of consciousness will necessarily have the capacity for the equivalent of acid trips. This seems doubtful. To start with, LSD is clearly a chemical matter and digital simulations of consciousness generally neglect the hugely complex chemistry of the brain in favour of the relatively tractable (but still unmanageably vast) network properties of the connectome. Of course it might be that a successful artificial consciousness would necessarily have to reproduce key aspects of the chemistry and hence necessarily offer scope for trips, but that seems far from certain. Think of headaches; I believe they generally arise from incidental properties of human beings - muscular tension, constriction of the sinuses, that sort of thing - I don't believe they're in any way essential to human cognition and I don't see why a robot would need them. Might not acid trips be the same, a chance by-product of details of the human body that don't have essential functional relevance?

The worst thing, though, is that Smart seems to have overlooked the main merit of the Turing Test; it's objective. OK, we may disagree over the quality of some chat-bot's conversational responses, but whether it fools a majority of people is something testable, at least in principle. How would we know whether a robot was really having an acid trip? Writing a chat-bot to sound as if were tripping seems far easier than the original test; but other than talking to it, how can we know what it's experiencing? Yes, if we could tell it was having intense trippy experiences, we could conclude it was conscious… but alas, we can't. That seems a fatal flaw.

Maybe we can ask tripbot whether it smells turpentine.

Cosmopsychism And The Blobject

12 December 2015



Is cosmopsychism the panpsychism we've all been waiting for? Itay Shani thinks so. While others start small and build up, he starts with the cosmos and works down. But he rejects the Blobject...

To begin at the beginning. Panpsychism is the belief that consciousness is everywhere; that it is in some sense a basic part of the world. Typically when people try to explain consciousness they start with the ingredients supplied by physics and try to build a mind out of them in a way which plausibly accounts for all the remarkable features of consciousness. Panpsychists just take awareness for granted, the way we often take matter or energy for granted; they take it to be primary, and this arguably gets them out of a very difficult explanatory task. There are a number of variants - panexperientialism, panentheism, and so on - which tend to be bracketed with panpsychism as similar considerations apply to all members of the family.

This kind of thinking has enjoyed quite a good level of popularity in recent years, perhaps a rising one. Regular readers may recall, though, that I'm not attracted by panpsychism. If stones have consciousness, we still have to explain how human consciousness comes to be different from what the stones have got. I suspect that that task is going to be just as difficult as explaining consciousness from scratch, so that adopting the panpsychist thesis leaves us worse off rather than better.

Shani, however, thinks some of the problems are easily dealt with; others he takes very seriously. He points out quite fairly that panpsychists are not bound to ascribe awareness to every entity at every level; they're OK just so long as there is, as it were, universal coverage at some level. Most panpsychists, as he rightly observes, tend to push the basic home of consciousness down to a micro level, which leaves us with the problem of how these simple micro-consciousnesses can come together to form a higher level one - or sometimes not form a higher one.

Thus combination issue is a difficult one that comes in many forms: Shani picks out particularly the questions of how micro-subjects can combine to form a macro-subject; how phenomenal experience can combine, and how the structure of experience can combine. Cutting to the chase, he finds the most difficult of the three to be the problems with subjects, and in particular he quotes an argument of Coleman's. This is, in brief, that distinct subjects require distinct points of view, but that in merging, points of view lose their identity. He mentions the simplified case of a subject that only sees red and one that only sees blue: the combined point of view includes both blue and red and the 'just-red' and 'just-blue' points of view are lost.

I think it requires a good deal more argumentation than Shani offers to make all this really convincing. He and Coleman, for example, take it as given that the combination of subjects must preserve the existence of the combined elements, more or less as the combination of hydrogen and oxygen to make water does not annihilate the component elements. Maybe that is the case, but the point seems very arguable.

Shani also seems to give way to Coleman without much of a fight, although there's plenty of scope for one. But after all these are highly complex issues and Shani only has so much space: moreover I'm inclined to go along with him because I agree that the combination problem is very bad; perhaps worse than Shani thinks.

It just seems intuitively very unlikely that two micro-minds can be combined. Two of the things that seem clearest about our own minds is that they combine terrific complexity with a strong overall unity; both of those factors seem to throw up problems for a merger. To me it seems that two minds are like two clocks: you cannot meaningfully merge them except by taking them apart into their basic components and putting something completely new together – which is no use at all for panpsychism.

For Shani, of course, combination must fail so that he can offer his cosmic solution as an alternative route to a viable panpsychism. He sets out his stall with the following postulates.

1. The cosmos as a whole is the only ontological ultimate there is, and it is conscious.
2. It is prior to its parts.
3. It is laterally dual in nature, having a concealed and a revealed side (the concealed side being phenomenal experience while the revealed side is the apparently objective world around us).
4. It is like a fluctuating ocean, with waves, ripples and vortices assuming temporary identity of their own.
5. The cosmic consciousness grounds the smaller consciousnesses within it.
6. Conscious entities are dynamic configurations within the cosmic whole.
7. These consciousnesses are severally related to particular surges or vortices of the cosmic consciousness and never fully separate from it.

That seems at least a vision we can entertain, but it immediately faces the challenge of the Blobject. This is the universal cosmic object championed by Terry Horgan & Matjaž Potr?. They are happy with the grand cosmic unity proposed by Shani but they go further; how can it have any parts? They believe the great cosmic consciousness is the Blobject; the only thing that truly exists; the idea that there are really other things is deluded.

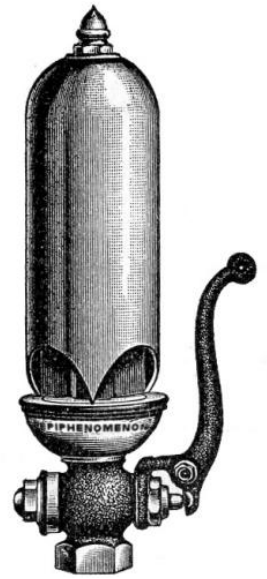
The austere ontology of the Blobject and its splendid parsimony can only be admired. We might talk more about it another time; but for now I'm inclined to agree with Shani that the task of reconciling it with actual experience is just too fraught with difficulty.

So does Shani succeed? He does, I think, set out, albeit briefly, a coherent and interesting view; but it does not have the advantages he supposes. He believes that starting at the top and working down avoids the difficult problems we encounter if we start at the bottom and work up. I think that is an illusion derive from the fact that the bottom-up approach has just been discussed more. I think in fact that just the same problems must recur whichever way we approach things.

Take the Coleman point. Coleman's objection is that in combining, two points of view lose their separate identity, while it needs to be preserved. But surely, if we take his blue-and-red pov and split it into just-blue and just-red we get a similar loss of the original identity. Now as I said, I'm not altogether sure that this need be a problem, but it seems to me clear that it doesn't really matter which way we move through the problem; and the same must be true of all arguments which relate different levels of panpsychist consciousness. Is there really any fundamental asymmetry that makes the top-down view stronger?

Implausible Materialism

20 December 2015



Physical determinism is implausible according to Richard Swinburne in the latest JCS; he cunningly attacks via epiphenomenalism.

Swinburne defines physical events as public and mental ones as private - we could argue about that, but as a bold, broad view it seems fair enough. Mental events may be phenomenal or intentional, but for current purposes the distinction isn't important. Physical determinism is defined as the view that each physical event is caused solely by other physical events; here again we might quibble, but the idea seems basically OK to be going on with.

Epiphenomenalism, then, is the view that while physical events may cause mental ones, mental ones never cause physical ones. Mental events are just, as they say, the whistle on the locomotive (though the much-quoted analogy is not exact: prolonged blowing of the whistle on a steam locomotive can adversely affect pressure and performance). Swinburne rightly describes epiphenomenalism as an implausible view (in my view, anyway - many people would disagree), but for him it is entailed by physical determinism, because physical events are only ever caused by other physical events. In his eyes, then, if he can prove that epiphenomenalism is wrong, he has also shown that physical determinism is ruled out. This is an unusual, perhaps even idiosyncratic perspective, but not illogical.

Swinburne offers some reasonable views about scientific justification, but what it comes down to is this; to know that epiphenomenalism is true we have to show that mental events cause no physical events; but that very fact would mean we could never register when they had occurred - so how would we prove it? In order to prove epiphenomenalism true, we must assume that what it says is false!

Swinburne takes it that epiphenomenalism means we could never speak of our private mental events - because our words would have to have been caused by the mental events, and ex hypothesi they don't cause physical events like speech. This isn't clearly the case - as I've mentioned before, we manage to speak of imaginary and non-existent things which clearly have no causal powers. Intentionality - meaning - is weirder and more powerful than Swinburne supposes.

He goes on to discuss the famous findings of Benjamin Libet, which seem to show that decisions are detectable in the brain before we are aware of having made them. These results point towards epiphenomenalism being true after all. Swinburne is not impressed; he sees no basic causal problem in the idea that a brain event precedes the mental event of the decision, which in turn precedes action. Here he seems to me to miss the point a bit, which is that if Libet is right, the mental experience of making a decision has no actual effect, since the action is already determined.

The big problem, though is that Swinburne never engages with the normal view; ie that in one way or another mental events have two aspects. A single brain event is at the same time a physical event which is part of the standard physical story, and a mental event in another explanatory realm. In one way this is unproblematic; we know that a mass of molecules may

also be a glob of biological structure, and an organism; we know that a pile of paper, a magnetised disc, or a reel of film may all also be “A Christmas Carol”. As Scrooge almost puts it, Marley’s ghost may be undigested gravy as well as a vision of the grave.

It would be useless to pretend there is no residual mystery about this, but it’s overwhelmingly how most people reconcile physical determinism with the mental world, so for Swinburne to ignore it is a serious weakness.